Spring 2019

Estimating Cognitive Workload in an Interactive Virtual Reality Environment Using Electrophysiological and Kinematic Activity

Christoph Tremmel
Old Dominion University, christoph.tremmel@gmx.de

ESTIMATING COGNITIVE WORKLOAD IN AN INTERACTIVE VIRTUAL REALITY ENVIRONMENT USING ELECTROPHYSIOLOGICAL AND KINEMATIC ACTIVITY

by

Christoph Tremmel
Diplom-Ingenieur December 2012, Karlsruhe Institute of Technoloy, Germany

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

BIOMEDICAL ENGINEERING

OLD DOMINION UNIVERSITY
May 2019

Approved by:

Dean J. Krusienski (Director)

Christian W. Zemlin (Member)

Krzysztof Rechowicz (Member)

Yusuke Yamani (Member)

ABSTRACT

ESTIMATING COGNITIVE WORKLOAD IN AN INTERACTIVE VIRTUAL REALITY ENVIRONMENT USING ELECTROPHYSIOLOGICAL AND KINEMATIC ACTIVITY

Christoph Tremmel
Old Dominion University, 2019
Director: Dr. Dean J. Krusienski

As virtual reality (VR) technology continues to gain prominence in commercial, educational, recreational and research applications, there is increasing interest in incorporating physiological sensors in VR devices for passive user-state monitoring to eventually increase the sense of immersion. By recording physiological signals such as the electroencephalogram (EEG), electromyography (EMG) or kinematic parameters during the use of a VR device, the user's interactions in the virtual environment could be adapted in real time based on the user's cognitive state. This dissertation evaluates the feasibility of passively monitoring cognitive workload via electrophysiological and kinematic activity while performing a classical n-back task in an interactive VR environment. The results indicate that scalp measurements of electrical activity and controller and headset tracking of kinematic activity can effectively discriminate three workload levels. Since motion and muscle tension can create co-varying task-related artifacts in EEG sensors mounted to the VR headset, decontamination algorithms were developed. The newly developed warp correlation filter (WCF) and linear regression denoising were applied on EEG, which could significantly decrease the influence of these artifacts. Analysis of the scalp recorded spectrum suggest two transient activity (termed pulse-decay effects) that impact feature extraction, modeling, and overall interpretation of workload estimation from scalp recordings. The best classification accuracy could be achieved by combining EMG, EEG and kinematic activity features using an artificial neural network (ANN).

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Dean J. Krusienski for the guidance and effort that he put towards the completion of this dissertation and throughout my Ph.D. career. I would also like to thank the members of the Committee Dr. Krzysztof Rechowicz, Dr. Yusuke Yamani, and the graduate program director, Dr. Christian Zemlin, for their consideration and participation in the thesis evaluation process. Finally, I would like to thank Morgan Mason for her assistance in the editing of this dissertation.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

In a world of rapid technological advances and increased everyday complexity, the assessment of mental workload is a very important factor to improve human machine interaction. The surge of portable electronic devices, such as mobile phones, laptops, GPS etc., in combination with increased global networking and the amount of available information has led to and increased multi-tasking with an overload of information in day-to-day life. In order to cope with this situation, the interdisciplinary research field "neuroergonomics" [1] originated and focuses on the relationship between the human brain and the performance for work and everyday tasks. One of the goals in this field is to monitor the mental workload level of human operators; as too much or too little workload leads to decreased performance, vigilance and increased security risks and therefore lowers productivity and reliability [2].

Consequently, it is important to monitor the workload level to keep the operator engaged without over or under loading. Paradoxically, this crucial engagement is not without difficulty. Self-monitoring leads to an even further increase in workload and is prone to misjudgment [3]. An ideal system could circumvent this concern by measuring and reacting to operator states; this action is termed "adaptive aiding" [4]. The possibilities for this technique are endless, as it could not only be applied in work environments, but also in education, interactive entertainment, sports, etc. In particular, video games and virtual reality applications could directly react to the operator's workload level and manipulate the environment in order to adjust to specific user needs.

There are three main categories of measures to assess workload in human subjects: subjective measures, performance-based measures and physiological measures [5]. Subjective and performance-based measures work well for evaluating single tasks or scenarios but do not work well for everyday situations since they require immediate feedback after a task. In contrast, physiological measures stay in the background, allowing application to multiple previously unidentified tasks and return information about the processes of the human body. These measures can then be interpreted by algorithms to draw inferences about the workload of users. While various peripheral physiological measures like pupil size, eye blinks, respiration, or heart rate have been evaluated, measures of the central nervous system like electroencephalography (EEG), functional near infrared spectroscopy (fNIRS) or electrocorticography (ECoG) are expected to produce more reliable results [6,7]. The most common measure in this field is EEG, and using EEG in a closed-loop control system is called "brain-computer-interface" (BCI). Researchers in the field of BCIs have focused their effort initially on the active brain control of systems like mouse cursors, communication devices or prosthetics, but lately more passive applications like workload estimation or mental state prediction have occurred [8].

1.2 BRAIN-COMPUTER-INTERFACES

A BCI is defined as a direct communication pathway between the central nervous system and an external device [9]. BCIs are made up of four main components: signal acquisition, feature extraction, feature translation and device control. The acquisition of a signal can either be performed invasively or non-invasively. While invasive techniques like ECoG or depth electrode recordings provide a higher spatial resolution and a better signal-to-noise ratio (SNR), the long-term viability and efficacy must be established to justify the surgical risk. Non-invasive techniques like EEG or fNIRS record the brain activity without penetrating the scalp, but suffer from a lower SNR

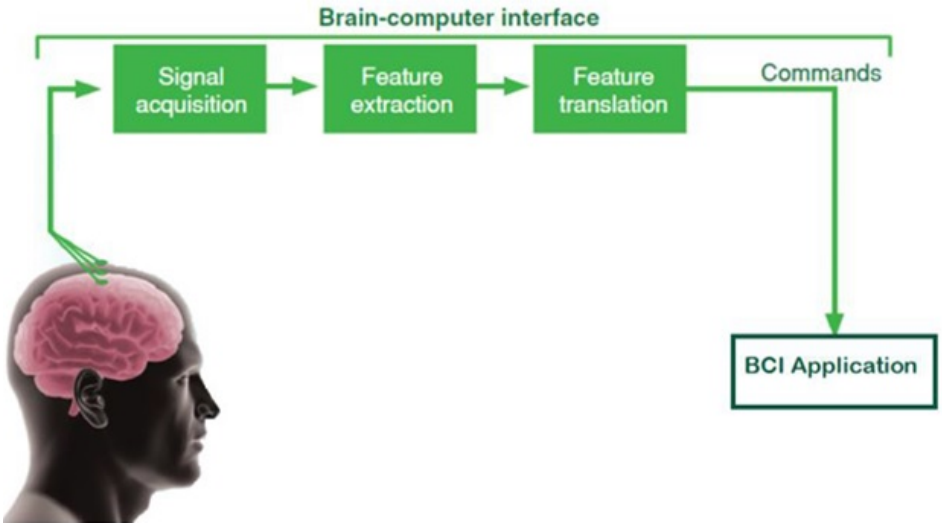

FIG. 1: An overview of the four BCI components [10].

and spatial resolution due to the distance and attenuation between electrode and brain tissue. After the acquisition, features will be extracted using signal processing algorithms like temporal and spatial filters, statistical decomposition methods or frequency spectrum analysis to decontaminate and decode the brain signals. In the third step, features will be translated to device commands using classification or modeling algorithms to predict the users intend. The translated features provide the input for external devices like robotic arms, cursors on a computer screen or wheel chairs.

1.3 MOTIVATION

With the recent surge of affordable, high-performance virtual reality (VR) headsets, there is unlimited potential for applications ranging from education, to training, to entertainment, to fitness, and beyond. As these interfaces continue to evolve, passive user-state monitoring can play a key role in expanding the immersive VR experience and tracking activity for user well-being. By recording physiological signals such as the EEG, EMG, and/or kinematic parameters during use of a VR device, the user's interactions in the virtual environment could be adapted in real time based on

the user's mental state such as cognitive workload level. User-state feedback to VR system for online modulation of task environments is critical for optimizing human performance in a myriad of professional environments including aviation, healthcare, defense, and driving, to name a few. The current generation of VR systems comes with high-precision tracking so that body movements in VR are synchronous with the physical movements. These systems also provide a logical, convenient, and unobtrusive framework for mounting EEG and EMG sensors. Additionally, recent advances in dry EEG electrodes and motion artifact suppression further increase the practicality of integrating electrophysiological measures into VR headsets. Passive monitoring of cognitive state is more practical than active control (e.g., VR environment navigation via mental imagery or eye movements) in terms of becoming a distraction to the user because it is less sensitive to noisy state estimates. This dissertation aims to demonstrate that estimates of cognitive workload based on EEG, EMG, and/or kinematic parameters can be reliably obtained during interactive use of VR.

1.4 OBJECTIVES AND APPROACH

In order to verify that EEG measures of cognitive workload can be reliably duplicated using an interactive VR environment, the well-established n-back task [11] has been adapted for modulating cognitive workload into an immersive virtual environment using a commercial VR headset (HTC VIVE). For the classical n-back task, participants are presented with a series of symbols and are asked to respond when the current symbol matched the symbol presented "n" symbols ago in the sequence. To adapt the task to a more immersive, game-like virtual environment, stimuli consisted of a series of colored balls presented on a virtual podium displayed through the VR headset. The participants perform the n-back task by physically reaching for the virtual balls and placing them into a target or non-target receptacle. This design is a deliberate attempt to move beyond controlled and sterile experimental

environments towards a more practical VR application, in which there is a detailed and potentially distracting environment, where the user is physically interacting with objects to perform a task. During the task performance, EEG was collected and analyzed offline to characterize the electrophysiological correlates of mental workload and develop an appropriate processing pipeline for predicting mental workload from the electrophysiological measurements.

The remaining chapters of this dissertation are organized as follows. Chapter 2 presents the background of this experiment highlighting the assessment of workload, the n-back task, which is used for creating different workload levels and the behavior of electrophysiological activity to increased workload. The current state of the art for BCIs in VR environments is also described. Chapter 3 covers the state of the art signal processing and machine learning algorithms that are used in workload estimation. The experiment used for this dissertation and the parameters for data acquisition and synchronization are detailed in Chapter 4. The first part of the analysis is presented in Chapter 5 and focuses on the evaluation of the new VR setting using traditional signal processing and machine learning methods. The frequency information of the established theta, alpha, and beta bands (4-30 Hz) were extracted from eight EEG channels and classified using an artificial neural network (ANN) and linear discriminant analysis (LDA) with a 5-fold cross validation. The result could validate the virtual reality setup, as it was comparable to traditional workload estimation experiments.

In Chapters 6 and 7, controller and headset movements as well as EMG data were integrated, since motion and muscle artifacts were contaminating EEG in the selected frequency range. Since both measures are also modulated by workload they were tested as features for classification and could achieve positive results. EEG was

decontaminated using linear regression denoising and the newly developed warp correlation filter (WCF), which significantly reduced the correlation between the contamination sources and EEG. By investigating standard score plots of the EEG features in Chapter 8, two effects could be identified that elevate EEG frequency features at the beginning of each trial and at the beginning of the experiment. These effects could also be verified in the standard score of the variance of controller movements. It could also be confirmed that while averaged EEG features behave as previous research has stated, individual features behave differently for each subject so that classification across subject or the acquisition of statistical significant features did not provide any useful results. Finally, all acquired data are combined in Chapter 9 to achieve the best possible classification accuracy. Chapter 10 concludes the dissertation by discussing the main contributions and possible future directions of this research.

CHAPTER 2

BACKGROUND

2.1 ESTIMATING WORKLOAD

Mental workload does not have a universal definition. Most accepted definitions describe workload as the relationship between the attentional and processing resources demanded by a task and the available resources supplied by an operator [12]. The exogenous task demands are factors outside the operator; these include task difficulty, task priority and situational contingencies. The endogenous supply of attentional resources are factors inside the operator; these include perceiving, updating memory, planning, decision making etc. and are modulated by the operator's skills and experiences [13]. Sub-optimal workload can mean either overload or underload and reduces the operator's performance. Workload cannot be measured directly but only indirectly through the measurement of related variables. Typically, the measures are divided into three main groups: performance-based measures, subjective measures and physiological measures [5]. Each category of measures has its own advantages and disadvantages; therefore, careful planning is required.

Performance-based measures can be split into two sub-categories, the primary task measures and the secondary task measures. Primary task measures assess the capability of the operator to perform the experiment's main task, whose workload is under investigation. The expectation is that speed and/or accuracy of the operator will decrease as the difficulty, and therefore the workload increases. Primary task measures have proven to be sensitive to changes in workload but perform unreliably for low levels of difficulty where the operator is able to compensate for increased

demand. Secondary task measures require an external task which is not part of the main setup. The performance of the secondary task is thereby the indicator of workload demand of the primary task though not all secondary tasks have proven sensitive to primary task manipulations [14]. Additionally, since the secondary task may interfere with the primary task, careful experiment design is required. Subjective measures make use of questionnaires or checklists such as the "NASA task load index scale" or the "modified Cooper Harper scale." While they can reliably assess the operator's workload, additional time is required to complete these questionnaires.

Finally, physiological measures represent all bio-signal measurements of the human body that indicate a change in workload. They can be divided into two subcategories, the peripheral measures and the central measures. The measures of the peripheral nervous system include cardiac (heart rate, heart rate variability), respiratory (respiratory rate, volume per breath), ocular (blink duration, eye blink rate, pupil size), blood pressure, dermal (electrodermal activity) or vocal measures (pitch, loudness). The measures of the central nervous system consist of EEG and functional near infrared spectroscopy (fNIRS). Each of these measures possesses its own advantages and disadvantages.

The most used physiological measures according to a review study were cardiac measures acquired by an electrocardiogram (ECG) [15]. The muscles of the heart create a stereotypical repetitive electrical potential that can be required using up to 12 ECG electrodes. The heart rate is defined as the number of contractions of the heart per minute, while the heart rate variability indicates the variation of the time between two contractions. Decomposing the time signal of the heart muscle in the frequency domain reveals three major frequency bands: low frequency (0.02 - 0.06 Hz), mid frequency (0.07 - 0.014 Hz) and high frequency (0.15 - 0.5 Hz) [16]. Generally, heart rate increases with increasing task demands [17]. It has been demonstrated that heart rate variability decreases with increased work load but eventually increases

gradually over long time scales, which may possibly be attributed to fatigue [18]. The mid frequency band has been shown to differentiate between low to intermediate task demands and is mostly used to distinguish between workload and rest [17].

The most common respiratory measure is the respiration rate. It can either be acquired from the ECG signal as the electrodes move due to the lung contractions or from tension measures using a respiration sensor belt. Once the respiration rate and the lung volume are known, the respiration volume can be calculated. The respiration rate generally increases [19] while the respiration volume decreases with rising task demands [20].

Electrodermal activity or galvanic skin response has shown to be sensitive to sudden workload changes but decreases over time [21]. To acquire the electrodermal activity, skin conductance is measured between two electrodes placed on the skin. The conductance is affected by the state of the sweat glands, which are controlled by the sympathetic nervous system.

Blood pressure is the pressure that circulating blood exerts on the walls of blood vessels. It is acquired by using a sphygmomanometer, an inflatable cuff, that is wrapped around an arm and measures the blood pressure of the artery. An increase in blood pressure is associated with increased task load [22] but the measure is heavily influenced by the participant's condition. Physical activity, sleep, digestion, stress etc. [23] are known to modulate blood pressure.

All ocular measures can be acquired using eye trackers. Most modern eye trackers use infrared light and record the reflection of that light off the cornea. The resulting reflection is then used to calculate the gaze direction. Blink rate and duration are shown to be especially influenced by task demands with a high visual component and decrease for higher demands [24]. Pupil size increases for increased task demands [25]. Likewise, vocal measures have also demonstrated modulation via workload. Speaking frequency (pitch) and vocal intensity (loudness) increased significantly with increased

task demands [26].

As EEG is the measure used in this dissertation it will be described in greater detail in the upcoming chapters. The other measure of the central nervous system, which is frequently used in workload estimation, is functional near-infrared spectroscopy. Neural activity in the brain is triggering a change in blood oxygenation. FNIRS is using this fact by irradiating the brain with near-infrared light (620 - 1000 nm) which disperses through most of brain tissue but is absorbed by hemoglobin. Hemoglobin is a molecule in red blood cells responsible for oxygen transportation. The levels of oxygenated and deoxygenated hemoglobin can be estimated from the resulting infrared light which gives information about the neural activity in the measured brain region. For increased task demand the levels of oxygenated hemoglobin increase while the levels of deoxygenated hemoglobin decrease [27].

2.2 N-BACK TASK

The n-back task is a sequential performance task which is popular within neuroimaging and assessment in cognitive neuroscience. Subjects are presented with a sequence of stimuli and have to indicate whether the current stimulus matches the stimulus "n" steps earlier. The task difficulty is controlled by the variable n and modulates the working memory load on the subject. \N" commonly ranges between 0 and 3, with the stimulus being fixed for $n = 0$. Common stimuli are letters, colors, numbers or locations on a board presented visually or auditorily. The basic structure of the n-back task using letters as stimuli for $n = [1, 2, 3]$ can be seen in Figure 2.

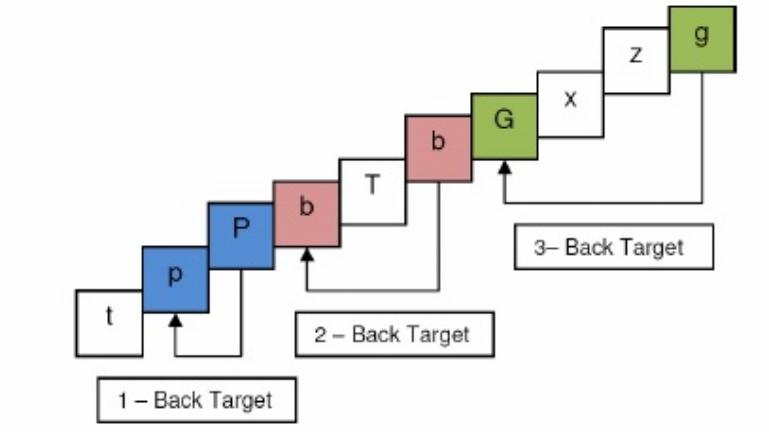

FIG. 2: Example of the n-back task using letters [28].

The performance of correctly identified stimuli drops from nearly 100% for 0-back to 75% for 3-back, the reaction time increases from 550 ms to 850 ms [29]. Even though the n-back task is a widespread measure of working memory, several studies found a low correlation between n-back and accepted working memory assessments. The performance of the n-back tasks seems to have a higher correlation with fluid intelligence [30] and speeded working memory tasks [31]. A meta-analysis of multiple neuroimaging study found evidence of six key frontal and parietal regions activated by the n-back task: bilateral and medial posterior parietal cortex, bilateral premotor cortex, dorsal cingulate/medial premotor cortex, bilateral rostral prefrontal cortex, bilateral dorsolateral prefrontal cortex and bilateral mid-ventrolateral prefrontal cortex [32].

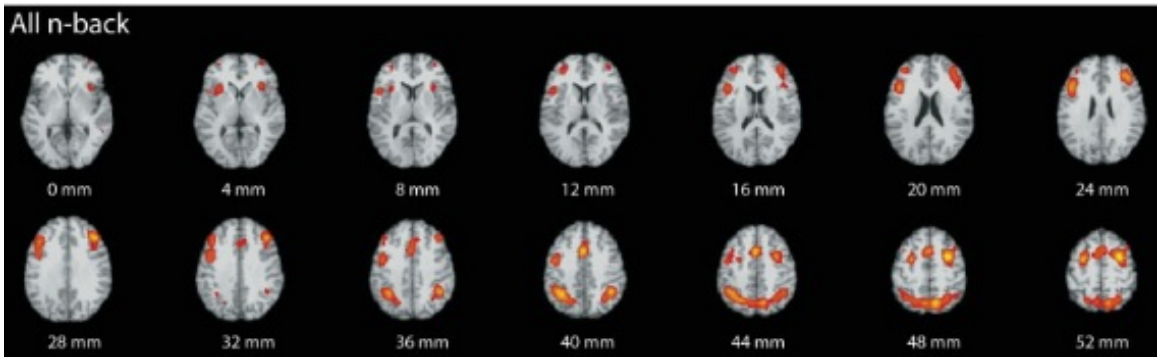

FIG. 3: Activated brain regions during the n-back task [32].

2.3 BCIS FOR WORKLOAD ESTIMATION

BCIs for EEG are classifiable in two categories, the voluntariness (passive vs. active) and stimulus dependency (independent/induced vs dependent/evoked) as it can be seen in Figure 4. Active BCIs are not suitable for workload estimation as it would require a control interface which would interfere with the workload task.

CLASSIFICATION	STIMULUS INDEPENDENT	STIMULUS DEPENDENT
ACTIVE	MOTOR IMAGERY	P300, SSVEP
PASSIVE	WORKLOAD ESTIMATION, EMOTION ESTIMATION	P300-BASED WORKLOAD ESTIMATION, ERROR RELATED POTENTIALS

FIG. 4: EEG paradigms classified into active/passive and induced/evoked.

There are two major approaches for workload estimation with EEG, passive stimulus dependent and passive stimulus independent BCIs. The passive stimulus dependent workload estimation for BCIs works by classifying event related responses. ERPs are brain responses to specific sensory, cognitive, or motor events and occur between 50 and 500 ms after the event. They all have a similar shape but differ slightly depending on the event as presented in Figure 5. Increasing the difficulty of work

load tasks decreases the amplitude of the brain responses [6,33]. Stimulus dependent BCIs require fixed events like an appearing question and cannot provide any estimates outside these events. The passive stimulus independent workload estimation uses

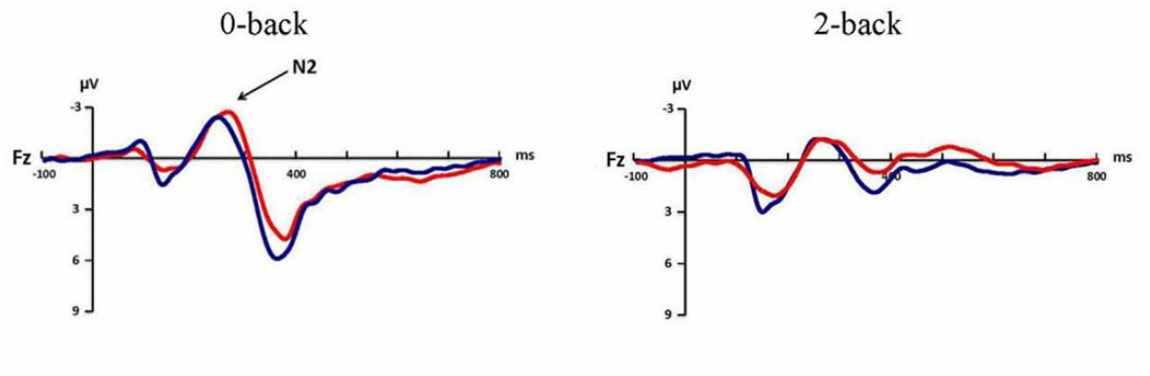

FIG. 5: Example waveform for 2 subject groups for 0-back and 2-back [34].

the frequency information of EEG responses to estimate workload. Increased task difficulty is increasing the energy in the theta band (4 - 8 Hz) while decreasing the energy in the alpha band (9 - 15 Hz) as shown in Figure 6 [33,35–37]. Stimulus independent BCIs can run in the background and update the estimates regularly without interfering with the tasks. While averaged results of most studies can confirm the frequency range, it is far more difficult to find these patterns for individual subjects. Multiple experiments report difficulties for analyzing individual activity caused by a large individual spread of patterns and highlight the importance of individually tuned classification models [33,35]. Some articles also report a positive correlation between the beta band (15 - 30 Hz) and increased workload [36,38].

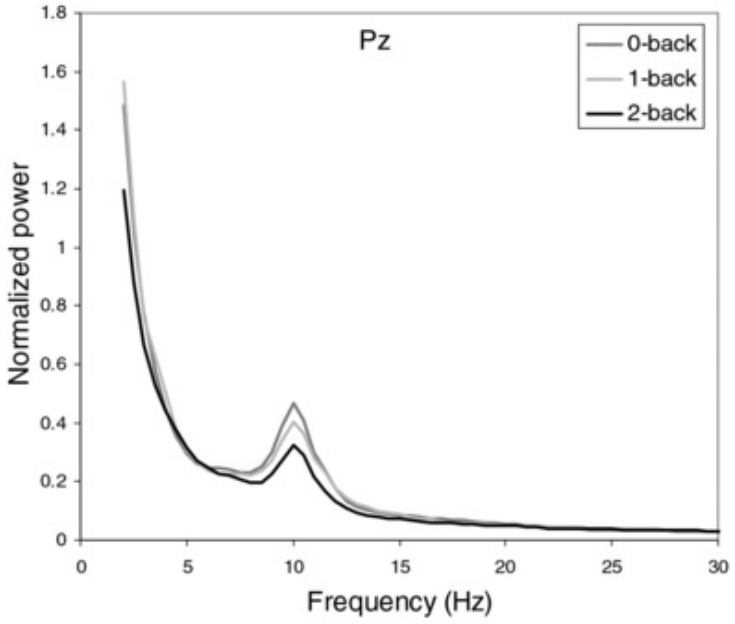

FIG. 6: Frequency spectrum for 0,1 and 2-back task [33].

Thereby, it is important to know that increased workload is accompanied with elevated muscle activity in several muscle regions including frontalis, temporalis, upper arm, shoulder, neck and back, even when subjects' posture was supported [39–41]. Muscle activity can be recorded using electromyography (EMG). The technique is detecting the electrical potential generated by activated muscle cells. EMG potentials range from 0 to 30 mV with a frequency from 0 to 300 Hz. Since the frequency range overlaps with EEG (0-60 Hz) signal contamination from cranial, facial and neck muscles is an unavoidable issue in experimental studies. Due to the difference in amplitude, up to 300 times bigger than EEG (0-100 µV), voluntary muscle activity can be detected easily. Involuntary muscle movement, such as facial expressions or muscle tension, however, produces far lower electrical activity. This electrical activity can be easily overlooked, yet still contaminates recordings. Figure 7 shows an example of EMG-free vs EMG-contaminated frequency spectra for three different electrode positions [42]. The EMG-free datasets were created by paralyzing the muscles in the head region of the subjects.

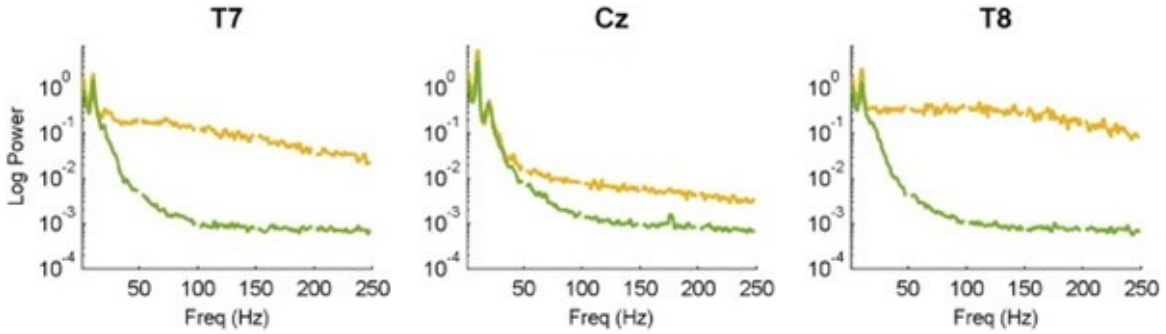

FIG. 7: EEG-free (yellow) and EMG-contaminated (green) EEG [42].

As shown, there is a significant difference in power in the frequency spectrum above 20 Hz and that this difference depends on the distance to the specific muscles (here temporalis muscle) as it is bigger for the electrode position Cz than T7 and T8, which are located at the temporal lobe. It is generally assumed that the power of muscle tension generated EMG artifacts will influence EEG mainly at about 20 Hz or higher. The popular solution is therefore to utilize a low-pass filter to exclude all higher frequencies, even though it neglects give any insights about the behavior of EEG for those frequencies. Additionally, previous workload estimation studies report two effects during these types of experiments. An initial transition effect for the first seconds of data for each trial has been reported as the human physiology does not adapt instantaneously with changing demands [43]. Also, there have been reports regarding an initial increase and slow decline in stress for workload tasks, allegedly caused by the laboratory setting [44].

2.4 BCIS IN VR

It has been found that VR immersion, presence, and task motivation can improve brain responses compared to traditional desktop-based feedback modalities. Contrary to the other domains, there has been relatively little work on passive biofeedback in VR. Two such studies used alpha-band power in the EEG to modulate either the controllability of the player's avatar in a video game or the transformation of the avatar into another physical form [45, 46]. Researchers at NASA Langley have also produced several unpublished demonstrations using similar video game paradigms (technology.nasa.gov/patent/LAR-TOPS-88), particularly targeted at using measures of EEG to encourage the user to maintain vigilance and composure during gameplay. Other passive neurofeedback studies successfully explored adaptation of the game difficulty based on changes in fNIRS signals [47], as well using changes in EEG to monitor items in the VR environment, that were detected by the user [48]. As EEG is prone to motion artifacts and VR technologies allow subjects to move and interact freely with its virtual environment the suppression of these artifacts is one of the most important challenges to be able to use the full potential of VR. Fortunately, the controller and headset tracker provide a more practical, simplified way to record and evaluate movements of the subject's upper body. There are also additional tracking sensors that can be applied to the legs and other body parts to achieve full-body capturing. This movement data can then be used to identify and eliminated motion induced contamination in EEG as previously achieved in gait related EEG research [49].

CHAPTER 3

SIGNAL PROCESSING FOR WORKLOAD ESTIMATION

3.1 SIGNAL ACQUISITION

Typical EEG amplifiers record the EEG signal between 128 - 512 Hz. The EEG signal's voltage is measured between electrodes, placed on the target region on the scalp and a ground electrode. A reference signal is measured between a reference electrode and the ground electrode then subtracted from every other signal in order to cancel out or at least reduce noise that is present on all channels (power line hum, heart beat, etc.). Reference and ground electrodes are usually placed at locations that are presumably inactive like the mastoid, the earlobe or the forehead to not induce additional noise onto the signals. It is imperative to keep in mind that the sampling rate always has to be twice as high as the expected frequency of the observed signal to satisfy the Nyquist frequency criterion. Undersampling of a signal can lead to aliasing, an effect that causes signals to take a different identity then they really are. In order to avoid aliasing, the sampling frequency of EEG signals has to be at least 120 Hz.

Electrodes are typically positioned according to the international 10-20 system which is based on the relationship between the electrode's location and the underlying brain area. "10-20" refers to the space between neighbor electrodes. The distances are either 10% or 20% of the total right-left or front-back distance of the skull. The letters of the electrodes are based upon the region of the brain they resemble ("F" for frontal lobe, "C" for central lobe etc.), the numbers refer to the placement to the left (uneven numbers) or right (even numbers) from the center line (marked with "z").

The electrodes for workload estimation are usual placed in the premotor, prefrontal and parietal brain areas as they show a positive linear relationship with increased workload in n-back tasks [32]. Bipolar recordings of symmetric electrode pairs on each hemisphere have also been featured in early experiments as differential asymmetry in the cerebral hemispheres can occur for workload estimation [50]. Measures of the human eye like blink duration, blink rate [51], pupil size, eye closure [52] etc., have been reported to be modulated by workload. Since artifacts from these movements can contaminate (mainly frontal) EEG channels, electrooculogram can be recorded in order to be able to identify these artifacts and either correct or discard recordings. Additionally, any number of peripheral measures can be added to increase the number of available information. Figure 8 depicts an electrode setup in the 10-20 system with reference and ground electrode that is connected to an amplifier.

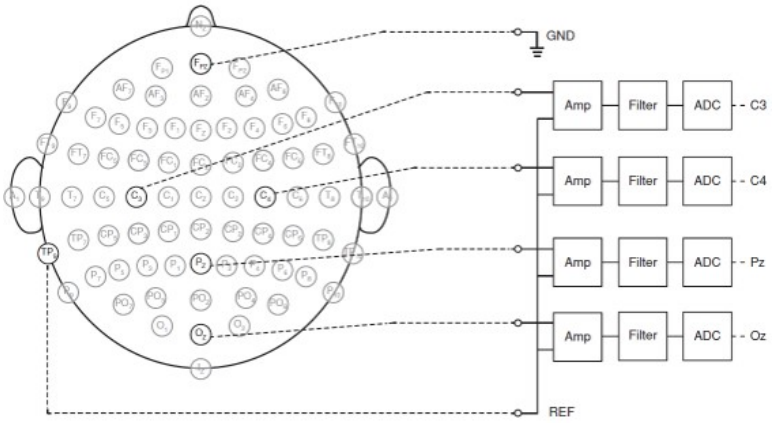

FIG. 8: Example electrode setup in the 10-20 System.

3.2 PRE-PROCESSING

The focused frequencies for workload estimation are the theta (4 - 8 Hz), alpha (8 - 14 Hz) and beta (14 - 30 Hz) bands. There are two approaches to isolate these frequency bands from EEG depending on which workload estimation technique is being used. When using the ERP responses in the time domain for classification, a

band pass filter has to be utilized in the selected frequency range. When using the spectral power features by transforming time windows in the frequency domain unwanted frequency bands can simply be disregarded by not including them as features for classification. Another possibility of feature generation is presented by Prinzel et al. [53] who developed an EEG-engagement index using the alpha, beta and theta band. It can be denoted as:

$$EEG = \frac{\beta}{\alpha + \theta}$$

The index was applied in a closed-loop system regulated task allocation and could improve the performance of subjects. In order to reduce noise that is present on multiple channels, spatial filter algorithms such as the Laplacian filter or common average reference (CAR) are often being used. Figure 9 shows an example on how create a spatial filter for the electrode position C3. The first setup portrays the electrode which is referenced to the ear lobe. In the second setup, CAR is computed by subtracting the average of all electrodes from C3. The small Laplacian filter subtracts the average surrounding electrodes while the large Laplacian filter selects the next closest surrounding electrodes for subtraction from C3.

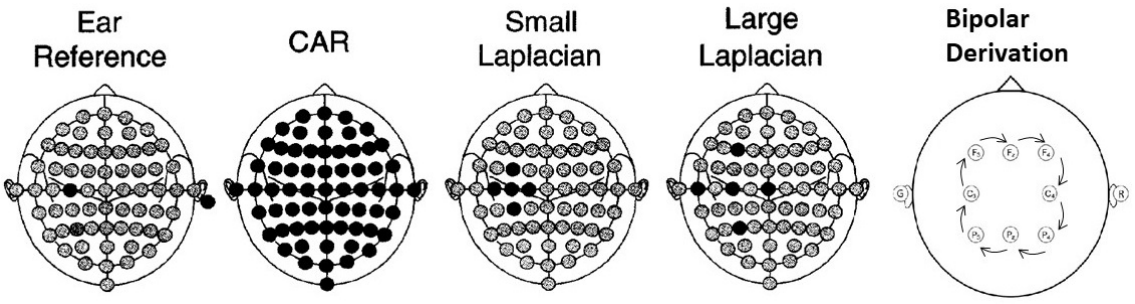

FIG. 9: Different spatial filter approaches adapted from [54].

McFarlan et al. [54] tested CAR, large and small Laplacian filter and compared the SNR to raw data that was referenced to the ear lobe. In an extended test of 26 - 86 hours per subject CAR and the large Laplacian filter could produce the better

results. Another spatial filter that is mostly utilized for a lower number of recording electrodes is the bipolar derivation [55] which subtracts the signal from one electrode to the one closest to it as displayed in Figure 9.

Since an EEG contains many different brain signals that might create noise for the required task, feature source separation algorithms are often used. Methods like principle component analysis (PCA) [56] or independent component analysis (ICA) [57] attempt to recover the desired signal from a combination of overlapping signals. In a cognitive state assessment competition, Kothe et al. developed a regression approach using ICA and PCA as pre-processing which in turn proved to outperform the other competitors [58].

PCA uses an orthogonal transformation to convert a dataset into a new coordinate system by computing a linear combination, the principle components, of the original variables. The principle components are then ordered by its variance as typically only the components with the greatest variance, are needed to describe a dataset. The transformed dataset makes it easier to identify patterns as it highlights the relationship between the variables of the dataset. In Figure 10, the three dimensional dataset is observed to be transformed into a two dimensional coordinate system without compromising much information.. PCA is often combined with ICA, as the representation of the data by principle components simplifies, and therefore accelerates, the work of ICA algorithms.

ICA decomposes a multivariate signal into components that are maximally independent and non-Gaussian. Maximum independence is thereby achieved through either minimization of mutual information or maximization of non-Gaussianity. ICA is often used as blind source separation to identify source signals from a mixed dataset without the need of additional information, such as the original signals or the mixing process. A typical example that explains ICA includes the “cocktail party problem,” a phenomenon of the human brain where a guest attending a cocktail party is able

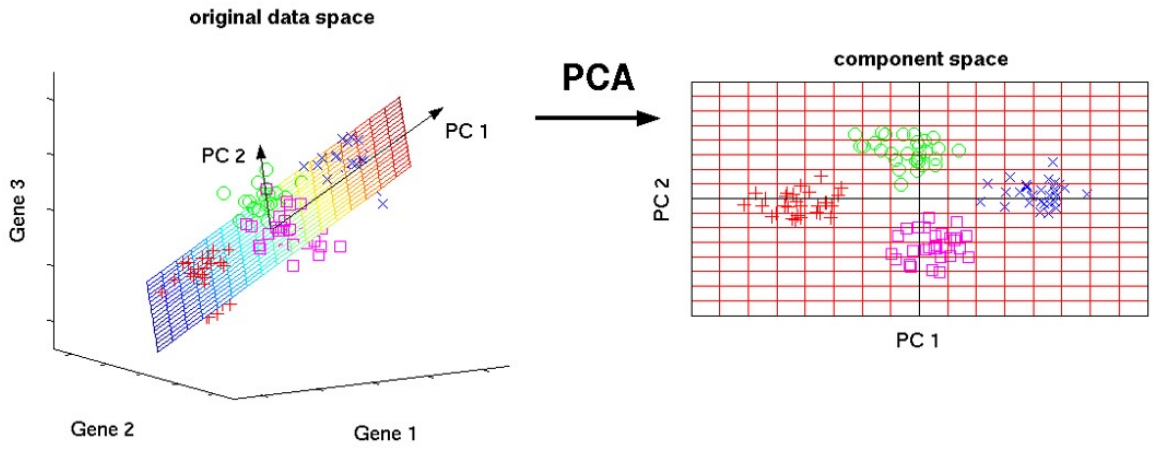

FIG. 10: Example how PCA works by remapping a three dimensional into a two dimensional dataset [59].

to focus his attention on a particular stimulus while ignoring the other stimuli. Figure 11 shows how the ICA algorithm can identify two sets of maximum independent components in one dataset.

3.3 FEATURE EXTRACTION

The passive stimulus dependent workload estimation needs to have fixed events in order to analyze the ERPs. The epochs for ERPs are normally from 0 to 1000 ms after the event. Down-sampling of epochs is common to reduce the number of samples and the amount of computing. The passive induced workload estimation can either use fixed events or continuous tasks. For fixed events the epochs are from -500 - 0 ms before or at the event to multiple seconds after the event. For continuous tasks the beginning and end of a difficulty level recording has to be cut to reduce transition effects and then remaining data is sampled into second or minute epochs, depending on the task. The window length has significant impact on the classification as longer windows produce better classification results [33]. The power band (or spectrum) features of the epochs are then created in the desired frequency bands using Fast

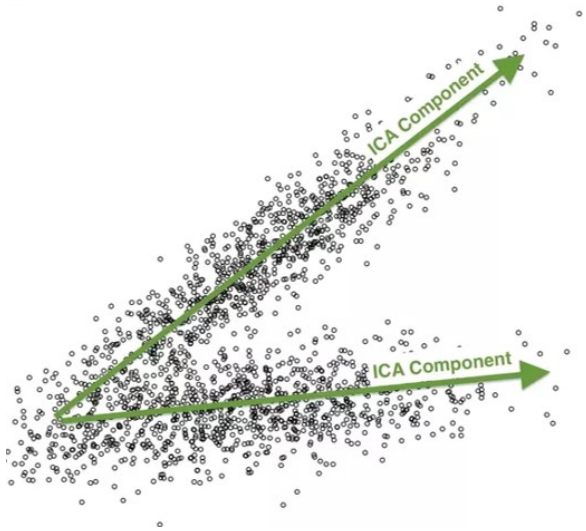

FIG. 11: Example how ICA identifies two sets of maximum independent components [60].

Fourier Transform (FFT) or Welch's method. The general Fourier Transform is a way of representing a time signal in the frequency domain by decomposing it in the frequencies it is made up. FFT is a computationally more efficient variant of the discrete Fourier Transform (DFT), the decomposition of discrete time signals. DFT transforms a sequence x_n of N numbers into its frequency representation X_k and can be denoted as:

$$X_k = \sum_{n=0}^{N-1} x_n * e^{\frac{-i*2*\pi}{N}kn}$$

The resulting representation consists of a phase value that shows the shift of a signal in time and the amplitude value that denotes the scaling of the amplitude. The power spectrum of the signal can then be estimated by squaring the amplitude and indicates the strength of the signal's variation in each frequency band. Welch's method is generating a frequency estimation by calculating the amplitude of overlapping time signal windows using FFT and then averaging these windows to reduce the variance of the estimate [61]. Other typical features that have been used in workload estimation are wavelet features, a correlation between wavelet function and EEG

signal for predefined number of resolution levels at each time position [62].

3.4 CLASSIFICATION

Classification algorithms are techniques that identify the category of new observations based on the category of already recorded observations. The already observed data is named training data, while the new observations are called testing data. Usually about 70% to 90% of the data [63] (depending on the feature parameters) is used for training and the rest for testing. To ensure that all observations are used for training and testing and to evaluate the whole dataset, cross-validation is often applied. Cross-validation rotates parts of training and testing data, as well as monitors the performance of the classifier. One of the most common cross-validation algorithms, "k-fold cross validation," divides the dataset into k parts and uses each of the k parts for testing while using the the other $k-1$ parts for training. The k results are then averaged to get the performance of the classifier.

Typical classification algorithms that are used for brain-computer-interface related research are linear classifier, like linear discriminant analysis (LDA) [64] or support vector machine (SVM) [65], non-linear artificial neural networks (ANN) [66], and variation of these. Linear classifiers categorize new observations based on a linear combination of the feature values. They are very popular for BCI applications. LDA is working similarly to PCA in trying to map a dataset into a different coordinate system as it can be seen in Figure 12. In contrast to PCA, which tries to create the principle components based on the variance, LDA tries to create new components that maximize the separability among these classes.

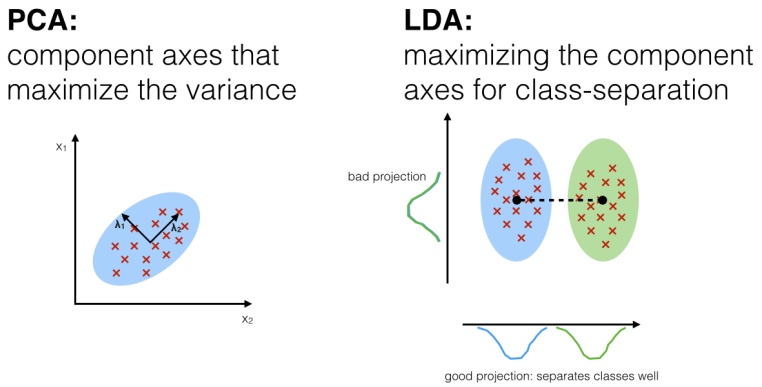

FIG. 12: Difference between PCA and LDA [67].

Maximizing the seperatibility can be denoted as:

$$max \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

With μ as the mean and σ as the variance of both classes 1 and 2. Once the two classes are distinguishable, LDA uses a hyperplane to seperate the two classes. For multi-class problems multiple hyperplanes are used. It is important to note that LDA assumes that the dataset is normal-distributed. Non-Gaussian datasets might result in a non-separable projection and will result in missclassifications. LDA is very popular as it requires low computational requirements, is simple to use and generally provides good results [68] A variation of LDA, called regularized Fisher's LDA (RFLDA), introduces a regularization parameter C that allows the penalization of training data classification errors. This prevents the classifier from generating the hyperplanes based on outliers.

SVM works similarly to LDA as it also uses a hyperplane to identify the classes. In contrast to LDA, the hyperplane is created by maximizing the margins; the distance between the nearest data points of both classes. A regularization term is used like for RFLDA to accommodate outliers on the training set. It is also possible to create a non-linear SVM by mapping the data into a higher dimensional space and applying a

kernel function. This is known as the kernel trick and is called Gaussion SVM. Figure 13 shows an example how to create a non-linear hyper plane in a higher dimensional space. The advantages of SVM consist of the good generalization properties of the margin maximization, and an insensitive behavior to overtraining and the curse-of-dimensionality [68].

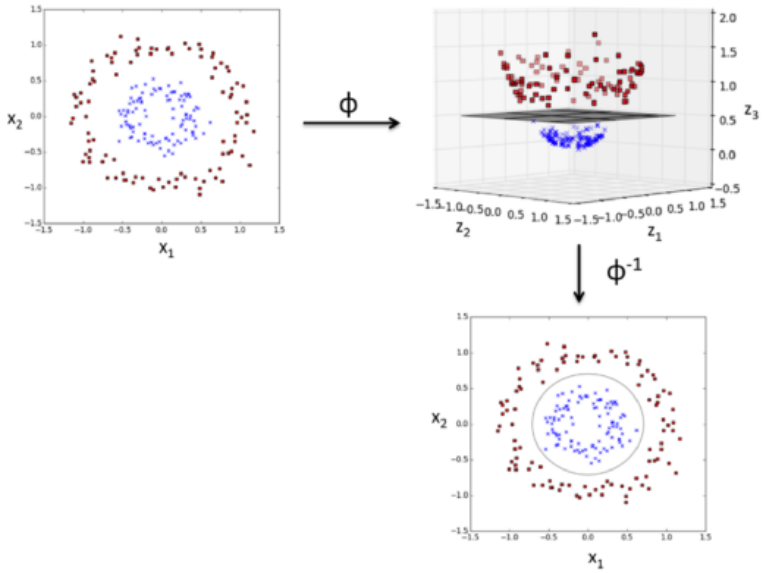

FIG. 13: Example of the kernel trick by creating a hyperplane in a higher dimensional space [69].

ANNs are machine learning systems that are inspired by biological neural networks, particulary the brain. It consists of a collection of connected nodes, which are loosely modeled after the central nervous system. These nodes are organized in layers: an input layer, an output layer and one or more hidden layers. Each layer can hold one or more nodes and each node from one layer is connected to all other nodes of the neighbor layer(s). An example of a basic feed forward ANN is shown in Figure 14. Each connection between two nodes represents a weight, so that the output is the combination of the weighted inputs.

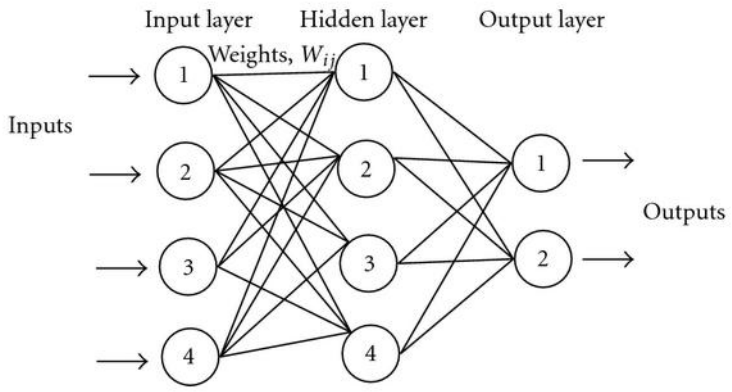

FIG. 14: Example of a fast forward neural network [70].

The formula of the output nodes denotes as:

$$y = W_3(i) \times \phi \left(\sum_{i=0}^{D} W_2(i) X(i) \right) + W_1(i) \times X$$

With W_1 as the weights of the input signal X, W_2 as the weights of the input to hidden layer $X(i)$ and W_3 as the weights for the output signal. σ is hereby the activation function that is applied before the output layer. Activation functions define the output of a node depending on its input. A binary step activation function for example would only give out the values 1 if the input is greater than 0 or 0 if the input is smaller than 0. One of the most used activation functions is the sigmoid function [71], it generates outputs from 0 to 1 as the input ranges from negative to positive infinity. ANNs are considered universal approximators, so that any continuous function can be approximated by an ANN of proper size. This makes them very flexible, but also sensitive to overtraining [68]. Each of these classifiers has different advantages or disadvantages, so a careful consideration is required.

Lastly, there is the group of feature dimensionality reduction methods that try to decrease the amount of data in order to reduce the processing time. Methods like PCA can reduce the number of samples by disregarding principle components with a low variance. Correlation-based feature selector (CFS) discards features with a low

correlation with the target variable.

3.5 EXAMPLES

Brouwer et al. [33] compared and combined ERP and spectral power features with the n-back task in three conditions (0-back, 1-back, 2-back). EEG and EOG signals were recorded from 35 participants. Letter stimuli were presented for 500 ms and two keys were used to select targets and non-targets in a time window of 2000 ms. An n-back difficulty block consisted of 48 letters and was repeated eight times. ERP features were extracted from 0 - 1000 ms after stimulus presentation and were downsampled from 256 to 100 Hz. Power band features were extracted from -500 to 1500 ms relative to the stimulus presentation and frequency bands from 2 to 20 Hz in steps of 0.5 Hz were selected. They did not find any significant classification result differences between the features and feature combinations with the combined features performing best at 67% in single trial classification. Increasing the number of trials for classification increases the classification accuracy but decreases the information transfer rate. The best results for two classes could be achieved by differentiating the 0-back and 2-back difficulty, which resulted in a classification rate of 88%. EOG features performed worse with 78%, a combination of both features lowered the classification accuracy compared to EEG to 84%. While the features in the alpha band generally seemed to contain the most relevant features, each subject showed an individual reaction to increased task difficulty which made generalized statements difficult.

Hogervorst et al. [6] used the n-back task again in three conditions (0-back, 1-back, 2-back) to test a wider variety of physiological and central measures, ERPs, spectral power, skin conductance, respiration, ECG and eye measure features with SVM and an elastic net classification model. The experiment parameter were similar to Brouwer et al. [33]. For single features the ERP and spectral power performed

best with an average classification rate of 86%. Pupil size and respiration frequency also performed significant with 75% and 68% respectively. Other measures performed worse between 50 and 60%. A combination of the same type of features resulted in a similar result. EEG features produced the best classification rate with 87%, eye features with 75% and respiration with 70%. ECG and skin conductance achieved 60%. Combination of recorded values from the same type resulted in no significant improvement. Combining different groups of features resulted only in a modest, non-significant improvement.

CHAPTER 4

PROGRAM AND EXPERIMENT DESIGN

4.1 PROGRAM DESIGN

The HTC VIVE hardware system consists of a motion-tracked headset display, two motion-tracked hand controllers, and two "lighthouse" base stations that are capable of providing 6 Degree of Freedom (6DOF) tracking. The controllers are visible for users in the virtual reality. The virtual environment of this experiment was designed in Unity [72] and consisted of a closed room resembling a science fiction design, a podium on which the colored balls would spawn, a vase-like non-target receptacle and a treasure chest-like target receptacle on each side of the podium. The details of the environment were intentionally designed to replicate a video game-like simulation in order to increase to increase the level of immersion and test whether such visual detail has any impact on task performance, in comparison to prior, less-distracting desktop-based studies. The virtual environment before and after the first ball spawn is presented in Figure 15.

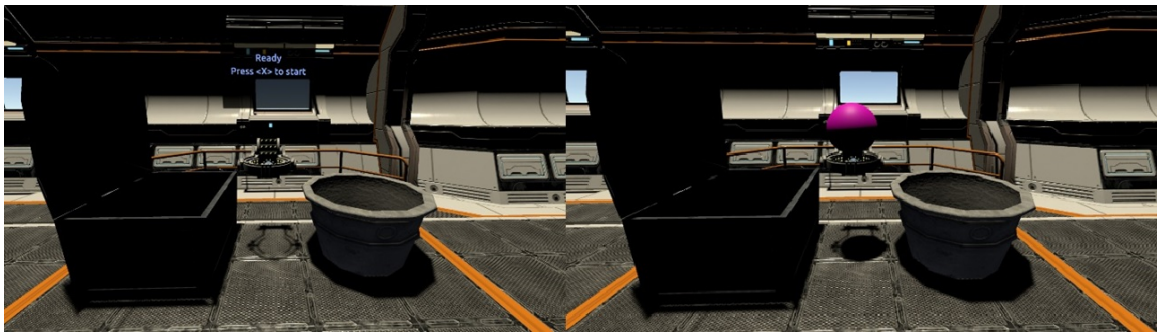

FIG. 15: The n-back task in the virtual environment before and after the first ball spawned.

Individual balls were presented directly in front of the participant and were colored yellow, red, blue, purple, or green according to McMilan at at. [73]. The different receptacles' designs were selected to help the subjects identify the correct one as the receptacles' positions were randomized across trials to avoid a movement bias. Participants could freely walk through the level but the interaction between player and environment was limited to the N-back task. Teleportation using the trackpad button of the controller was allowed in close proximity to the podium to ease positioning adjustments and was disabled after the start of the task. Above the podium a text field displayed the message "Ready. Press <X> to start.". Once the subject was prepared, the experimenter would press the X key and a three seconds countdown would indicate the start of the experiment. After the countdown, a randomly colored ball would spawn on the podium and an acoustic tone would indicate this. The subject then had a fixed time window to decide whether this ball was a target by putting it in the treasure chest or a non-target by putting it in the vase receptacle. A ball was grabbed by moving the controllers closer to it while pressing and holding the trigger button on the underside of the controller. Releasing the button would release the grip and the ball would drop on the floor. Placing the ball into one of the receptacles would make the ball disappear; failure to do so in the predefined time limit would result in a slow fading black screen and the restart of the experiment.

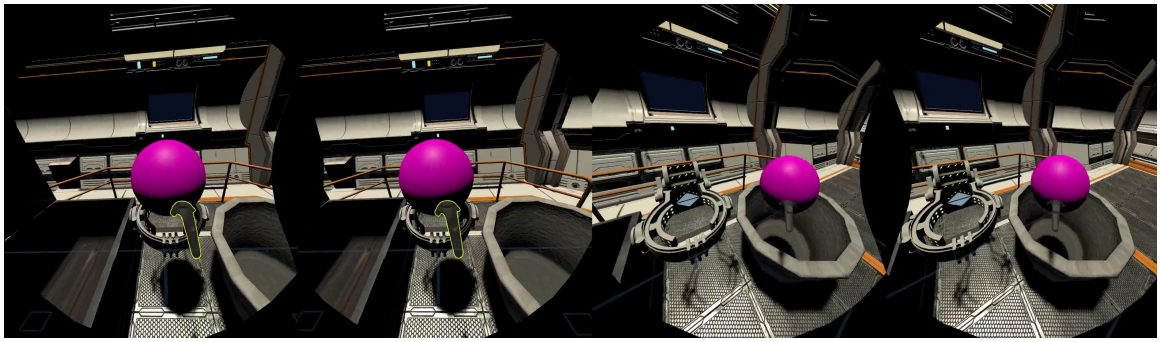

FIG. 16: The n-back task in the virtual environment using binocular view from the user's perspective before grabbing a ball and while putting it into the receptacle.

Figure 16 indicates an example from the participant's perspective while a ball was placed in the non-target receptacle. At the end of a trial the performance and the percentage of correctly placed balls were displayed in the text field above the podium to help engage participants as depicted in Figure 17. Each correctly placed ball gave one performance point; each incorrectly placed ball resulted in two negative points.

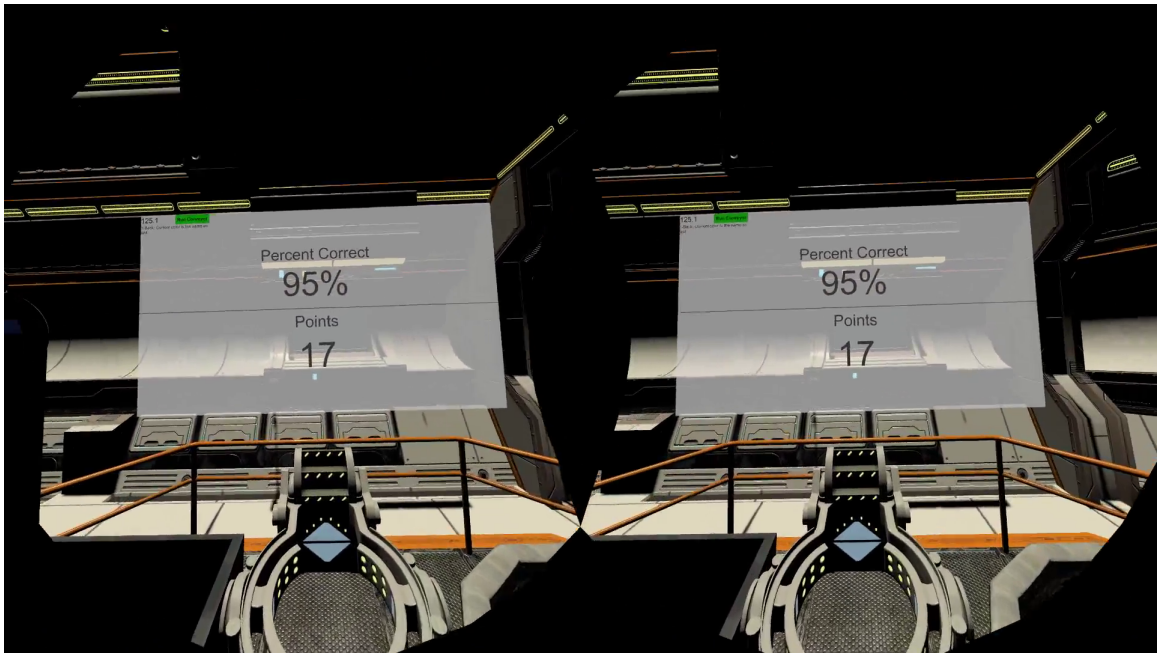

FIG. 17: The performance evaluation of the n-back task in the virtual environment using binocular view from the user's perspective after finishing a trial.

While the experiment was running, the program would record the HTC Vive controller's and headset's position, the color of the spawned balls, the item status indicating which balls are targets or non-targets, the receptacle in which the balls were dropped into, the start of the experiment after the countdown, the percentage of correct placed balls after the experiment and the time for each of these events. All the data were sent to the "AppConnector" of BCI2000 [74] using user datagram protocol (UDP) with an update rate of 32 Hz. BCI2000 would then synchronize the acquired data with the information from the 8-channel wireless biosignal amplifier (g.MOBIlab, Guger Technologies). The amplifier was connected to the PC via Bluetooth and was

recorded with an update rate of 256 Hz. The variable that tracked the start of each experiment would also be the trigger for BCI2000 to start recording and synchronizing both data streams.

The parameters of each experiment could be set using a text file. The parameter consisted of the IP address, the client and server port for the UDP connection, the experiment's name, the location of the target receptacle (left/right), the time interval in which a ball had to be dropped into a receptacle, the difficulty of the n-back task, a random seed number for the generation of the balls' color and the number of balls that would spawn. Additionally, there was an optional parameter to generate a minimum number of n-back targets and how much the minimum should be. This was added to always have a guaranteed number of targets.

4.2 EXPERIMENT DESIGN

Fifteen participants, ages 18-35 with a mean of 24.73 years (4 female), were recruited to participate in this experiment. Before the actual experiment, participants completed a visual acuity [75] test, followed by Motion Sickness Susceptibility Questionnaire short-form (MSSQ) [76] and Ishihara Color Blindness test [77]. The acuity test is using a Snellen eye chart that is placed in front of the subject. The subject has to identify letters that are reduced in size with each line. The visual acuity is then reported by the test distance divided by the letter size. The limit was set to be 20/40 and all participants scored a value of 20/30 or higher. The MSSQ short is an efficient form of the original MSSQ. It considers the past experiences and the reactions of participants to travel using different means of transportation.If a subject scored a 19 or above, the participant would be excluded from the experiment. However, since all subjects scored under the threshold, none of the participants were excluded. In the Ishihara color blindness test participants had to identify numbers and lines on colored plates, called Ishihara plates. Two subjects could not give correct answers to

all of the plates, but this had no impact on the performance of the n-back task.

After the screening process, participants were placed in a standing position on an anti-static mat approximately one meter in front of the recording computer, within the VR workspace. The EEG cap was mounted, and the amplifier and pre-amplifier were stored in a small shoulder bag. A processing bag was placed on top of the EEG cap to prevent gel contamination and the VR headset was positioned on it as presented in Figure 18. EEG was recorded using an 8-channel active electrode cap (g.LADYBIRD, Guger Technologies) having electrodes positioned based on the international 10-20 system. Specifically, electrode positions F3, Fz, F4, C3, C4, P3, Pz, P4 were used, which were selected based on neural activation from prior EEG and functional magnetic resonance imaging (fMRI) studies. EEG was collected using an 8-channel wireless biosignal amplifier, grounded and referenced to inked earlobes, and digitized at a 256 Hz.

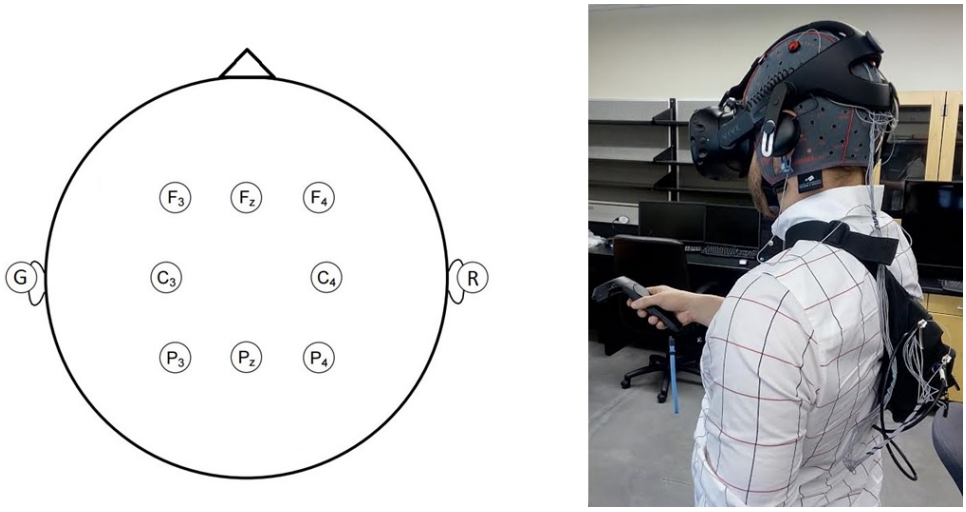

FIG. 18: The electrode configuration (left) and the experimental equipment on a participant, excluding the processing cap (right).

Participants were instructed to keep head movements to a minimum during the

trials and then completed a 1-back practice block until they could complete it without mistakes, to familiarize themselves with the VR system and the n-back task. Following the practice block, participants completed a series of three experimental blocks: 0-back, 1-back, and 2-back blocks with four trials each. Since most of the participants were unfamiliar with VR technology, and to prevent information overload, it was decided to not include the 3-back task and to keep a block-wise instead of a randomized structure of the difficulty levels. Each experimental trial consisted of a random sequence of 20 balls comparable to the number of stimuli from previous studies [73]. Each ball remained visible for four seconds, which is slightly longer than the typical two or three second windows that literature suggests [73]. However, since the VR n-back task is more complex than classical approaches, a shorter time would lead to an increased number of missed trials. The order of the experimental blocks was pseudorandom, and the order of the balls were random for each participant. It was assured that each trial consisted of at least two target stimuli using the function in the environment software. The task duration added up to $20 * 4 * 4 * 3/60 = 16$ minutes, the total duration of the whole experiment was designed not to exceed 25 minutes to reduce the risk of simulator sickness.

CHAPTER 5

ESTIMATING COGNITIVE WORKLOAD USING EEG

5.1 METHODOLOGY

Since only eight electrodes were recorded, a bipolar derivation was applied instead of CAR or a Laplacian filter, which are generally designed to be used with a larger number of electrodes. In order to avoid transition artifacts, excluding up to the first six windows has been empirically evaluated. Due to technical difficulties with the recordings at the end of a trial, the last ball presentation windows were excluded from further analysis. To reduce the dimensionality of the feature space, only the top components from PCA were retained and processed using ICA. The fastICA algorithm was used for this classification because of its reduced computation time compared to other ICA algorithms. The kurtosis was used as a measure of non-Gaussianity for the optimization criterion. The remaining data were segmented by ball presentation intervals of 0.5 to 4 seconds EEG.

The frequency spectrum of the data was then calculated using Welch's method and tested with 64 up to 1024 discrete Fourier transform points, sample window size of half the Fourier points and 60% overlap. For each segment, standard EEG spectral bands (theta (4-7 Hz), alpha (8-14 Hz), and beta (15-30 Hz)) were computed as the average of the spectral bins in each respective range. Since headset and controller movements produced artifacts that overlapped with the delta band, the Pearson correlation coefficient between each EEG channels and the movements was computed to justify the elimination of that range from further analysis.

To reduce occasional impulse-like artifacts due to sudden movements of the subjects like sneezing or losing their balance or due to the wireless transmission artifacts,

a conservative Hampel outlier filter was applied on the frequency bands. If one frequency feature would be nine times greater than the standard deviation of that frequency band over the course of the experiment, it would be corrected to the median of that frequency band. This was done to ensure that the classifier would not be trained with extreme values. The data were parsed by n-back cognitive workload level and the segments consisting of 28 data points per subject and observation were classified using linear discriminant analysis and an ANN (1 hidden layer, 100 hidden nodes) with a 5-fold cross-validation. The classification results are presented in boxplots. Boxplots show the median value in red and the 25 and 75 percentile as the beginning and the end of the box. The whiskers show the minimum and maximum value; crosses indicate outlier.

5.2 DELTA BAND ELIMINATION

A 4 Hz highpass filter with order 10 was applied onto the EEG data using zero-phase digital filtering, which does not distort the phase of the resulting signal as normal filtering would do. In Figure 19, boxplots show the Pearson correlation coefficients between the resultant distance of the controller and each EEG channel averaged over all subjects for the raw data, as well as the 4 Hz highpass filtered recordings. The analysis reveals that there is a correlation between the signals that has an average value of 7% for the frontal and central regions and 3% for the parietal regions with some subjects show even greater correlation up to over 20% when delta waves are included. The correlation values drop significantly ($p<0.001$) to 0% for all channels when the delta band is excluded. The high correlation results prove that there is a high contamination of movement artifacts in low frequency EEG. While the kinematic activity would likely increase the classification accuracy, as it is also modulated by workload, this would confound the objective of classifying exclusively using EEG. Moreover, using the highpass filter results in near-zero correlation and

therefore effective elimination of movement artifacts in EEG.

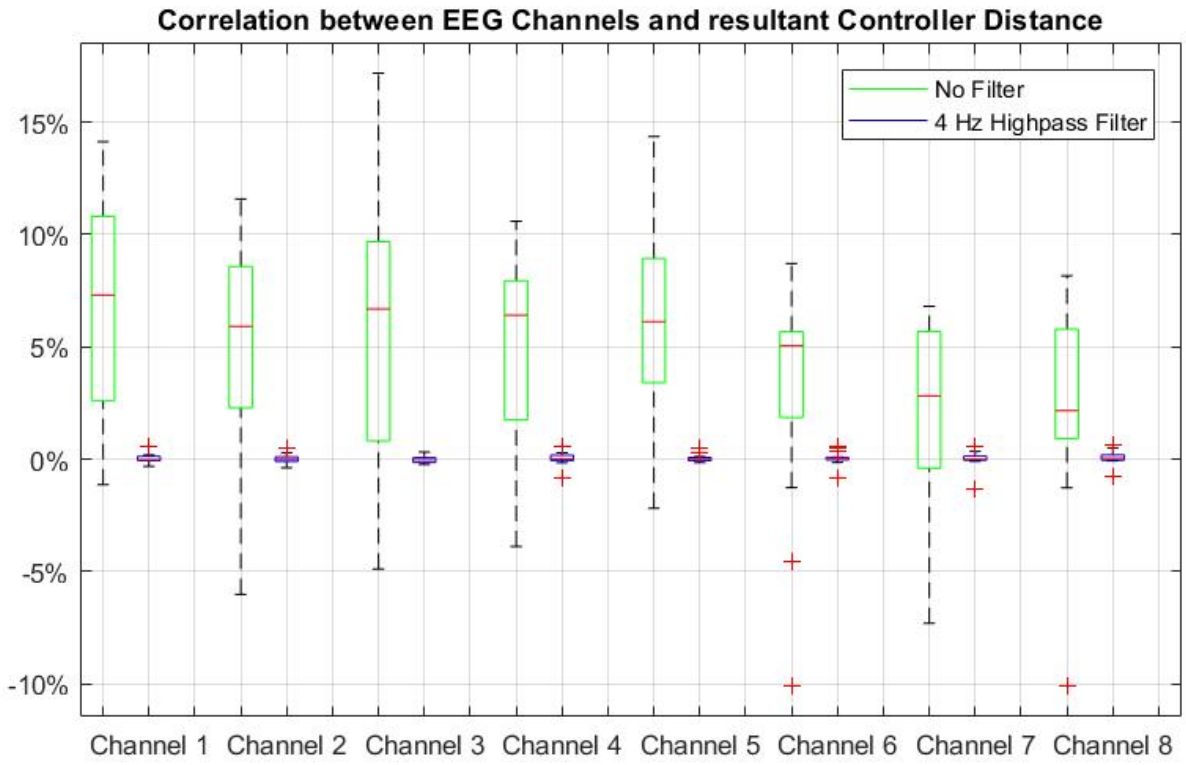

FIG. 19: Pearson correlation coefficient between EEG channels and resultant controller distance with highpass filter and without.

5.3 PARAMETER TESTING

While the signal processing and machine learning approach was selected based on several workload estimation papers [6, 33, 37], the individual parameters were empirically chosen. Table 1 shows the median and the variance of the output of the PCA algorithm averaged over each subject. It can be seen that the 8th PCA channel's variance is substantially smaller than the other channels', which indicates that the 8-dimensional EEG data can be remapped into a 7-dimensional coordinate system without losing much information.

PCA Channel	1	2	3	4	5	6	7	8
Median (µV)	0.65	-0.03	-0.058	0.02	0.01	-0.01	0.002	$-1.5 * 10^{-6}$
Variance (µV)	275.1	96.91	32.54	18.73	13.81	8.55	5.75	$2.3 * 10^{-5}$

TABLE 1: Median and variance voltage levels of PCA output channels.

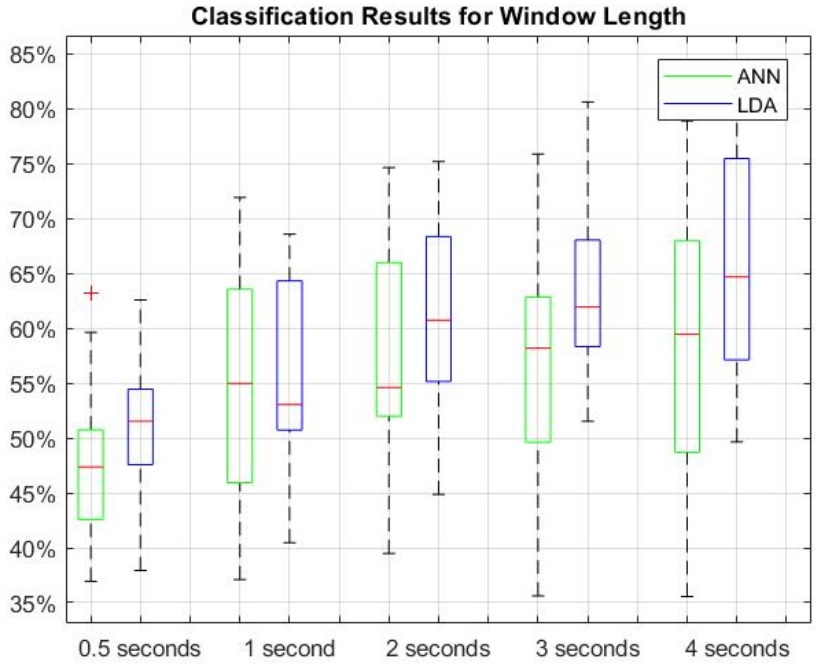

FIG. 20: Classification results for ANN (green) and LDA (blue) for different window lengths.

Figure 20 presents the classification results for window sizes of 0.5, 1, 2, 3 and 4 seconds. 4 seconds is the longest possible time for each window. The shortest window of 0.5 seconds is even shorter than the average reaction time 0.62 ± 0.09 seconds of the subjects to the stimulus. The reaction time was calculated by computing the movement speed, the first derivative of the positioning data. When the movement speed exceeded 10% of the maximum movement speed it was considered the movement onset. The time between stimulus presentation and movement onset was regarded as reaction time. The results show an increase of the classification accuracy for increased

window size. Starting with 50.1% for LDA and 48.3% for ANN for 0.5 second windows and ending with the best result at the longest window with 59% for ANN and 66% for LDA. This confirms findings in previous workload experiments that the window length significantly impacts the classification accuracy.

Literature in workload estimation research also reports a transition period at the start of each task sequence [43]. While the last window of each trial had to be excluded due to technical difficulties, Figure 21 shows the average classification results for removing none up to the first six windows from the analysis. From the boxplots it can be seen that excluding one to three windows generally has a positive, though not significant, effect on the classification accuracy of LDA and ANN. Excluding more windows results in a worse classification rate, as less training data is available for the classifier. Excluding no windows results in a classification accuracy of 64.7% for LDA and 58.3% for ANN in average. The maximum classification rate could be achieved for excluding the first three window. Since the difference between one and three windows is minimal, 67.7% vs 67.3% LDA and 60.2% vs 59.7% in average it was decided to only exclude one window in order to not further reduce the training and testing data for the classification.

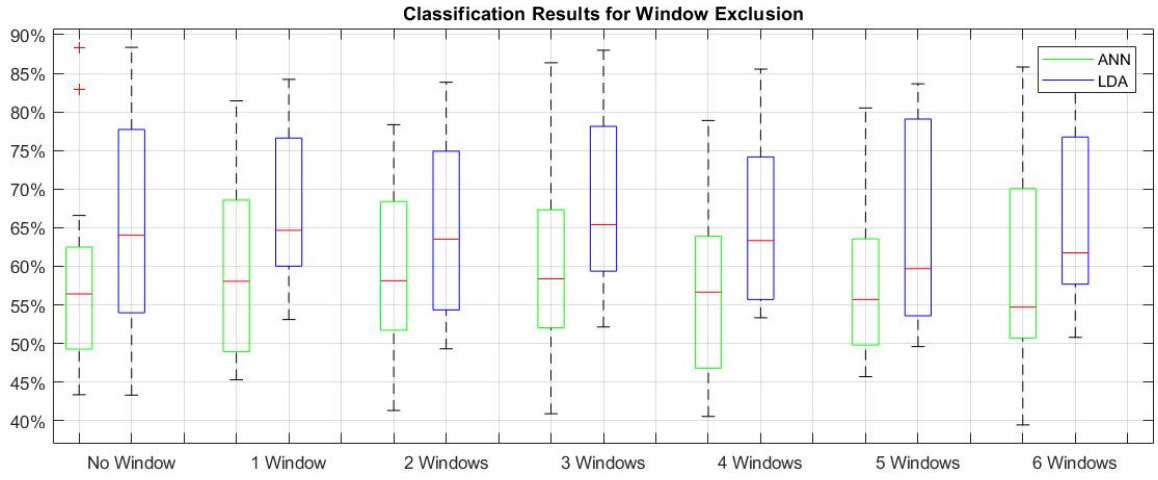

FIG. 21: Classification results for ANN (green) and LDA (blue) for excluding the first 0 - 6 windows.

Figure 22 shows the classification results for evaluating the number of fast Fourier points that have been used in Welch's power estimate algorithm. More FFT points generally result in a higher frequency resolution. 64 FFT points result in a 4 Hz, while 1024 points result in a 0.25 Hz resolution. It can be seen that the maximum accuracy that could be achieved for LDA is at 64 points with a 66.8% average while for ANN it is at 128 points with a 59.5% average. Increasing the number of points further will decrease the results.

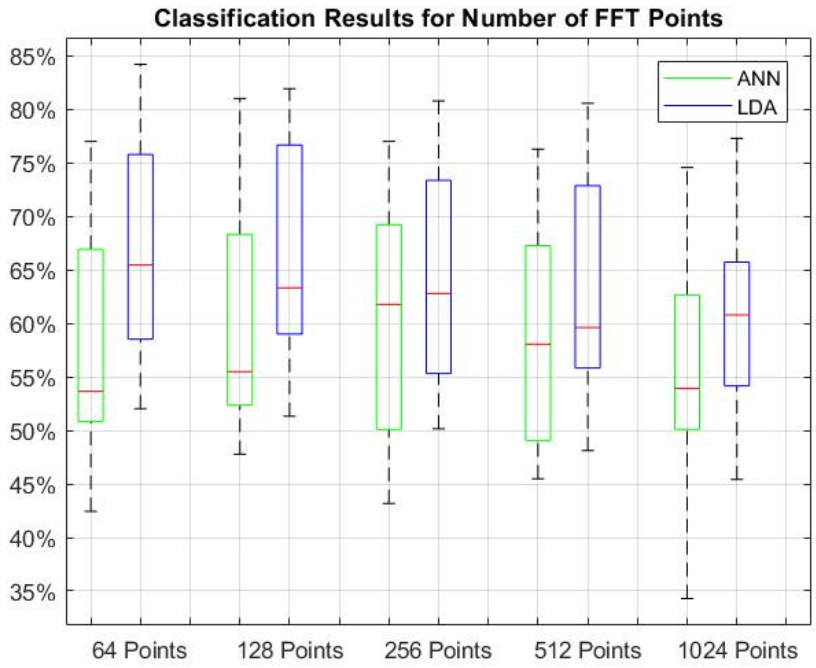

FIG. 22: Classification results for ANN (green) and LDA (blue) for different points of FFT.

5.4 RESULTS

The average task performance (correct bin placement) was 99.67±1.56% for $n = 0$; $98.17 \pm 4.5\%$ for $n = 1$; and $95.83\% \pm 7.49\%$ for $n = 2$. Fourteen participants scored above 80% on all runs; the remaining participant scored above 70% on all runs. Twelve of the participants scored above 90% on all runs. Figure 23 shows the results

for the best setup for an LDA classifier with a classification accuracy of 66.3% for three classes. The confusion matrix shows that the misclassification between class one and three was the lowest with 4.2% and 4.5%. Misclassification between class one and two or class two and three were similar between 5.2% and 6.8%. Class three has the best classification rate with 70.8%, while class two has the worst accuracy with 60%. The boxplot results for two classes show similar results. Therefore, the classification between 0-back and 2-back with 79.1% and was better than 0-back and 1-back with 76% and 1-back and 2-back with and 77.9% in average .

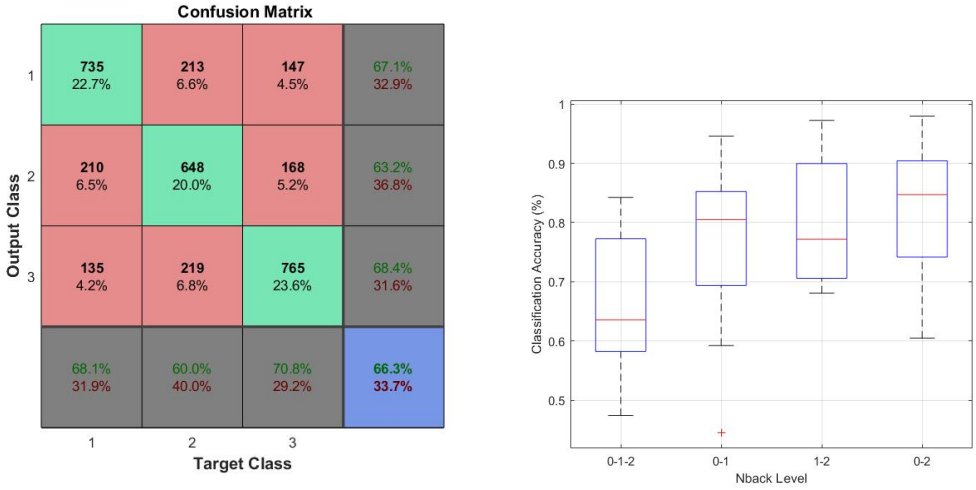

FIG. 23: Confusion matrix for 3 classes and boxplot for 3 and 2 classes with LDA classifier.

Using the same setup with ANN produces slightly lower results with an accuracy of 59.6% as shown in Figure 24. The confusion matrix reveals the same misclassification rate for all classes between 5.9% and 7.4%. The best classification rate was again class three with 62% before class one and two with 59.7% and 59.9%. For two classes the best results could be achieved for 0-back vs 2-back with 77.7%. Classification for 0-back and 1-back and 1-back and 2-back resulted in 73.4% and 74.7%.

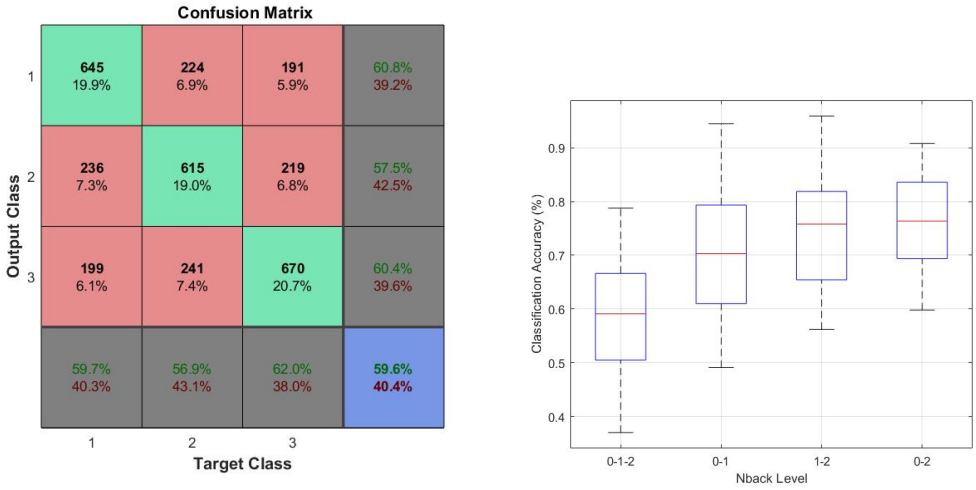

FIG. 24: Confusion matrix for 3 classes and boxplot for 3 and 2 classes with ANN classifier.

5.5 DISCUSSION

These results demonstrate that EEG-based estimates of cognitive workload can reliably be determined over short time intervals during the use of an interactive room-scale VR, therefore promoting further investigation of the potential of EEG as a tracking and feedback mechanism of cognitive states in VR. Using traditional frequency features from 4 – 30 Hz with a linear classifier, the difficulty level for 3 classes could be predicted with 66.3% accuracy, and with 79.1% for discriminating between 0-back and 2-back. Reducing the window size to 500 ms to exclude EEG data containing kinematic activity yielded in 50.1% accuracy. Using a non-linear classifier produced slightly lower results with 59.6% for 3 classes and 77.7% for discriminating 0-back and 2-back which suggests a more linear relationship between workload and EEG features of the theta, alpha and beta band as it has been reported by several studies.

It can also be seen that classification between 0-back and 2-back clearly produces better results, which is expected because of the greater difference between the two extreme workload levels. It has also been shown that there is movement contamination

in EEG due to high correlation between controller movement and low frequency EEG. A highpass filter was deployed to remove this contamination and mitigate this this correlation. Overall, when analyzing the parameters it was determined that longer windows, fewer FFT points and excluding the first window, as well as up to three, had a positive effect on the classification accuracy.

CHAPTER 6

ESTIMATING COGNITIVE WORKLOAD USING CONTROLLER AND HEADSET MOVEMENTS

The reaction time of subjects is a primary task measure in workload experiments and normally increases for increased task difficulty. This indicates that the kinematic activity should be delayed for higher difficulty of the n-back task and that it should be possible to estimate workload from the reaction time measured from movement onset. In order to validate this, the recorded movement windows will be used as features for classification. First, a Hampel outlier filter was applied onto all movement recordings. If during the recording the the controller or headset lost the connection to the lighthouse stations, the recorded value would be a large value outside of the range of the regular movement data. The Hampel filter compares all recorded values of each window to 20 times the standard deviation of that window. If a value is greater than this threshold it will be corrected to the median value of the window. This ensures that the classifier is not trained by extreme values produced by disconnections. Next, the features were downsampled from 256 to 32 Hz in order to return the features to their original sampling rate before the synchronization.

The following features were tested in four-second windows: controller movement in x, y and z-axis, resultant distance of the controller movements, headset movement in x, y and z-axis, resultant distance of the headset movements, combination of all headset movements, combination of all controller movements and combination of all movements. The resultant distance can be denoted as follows:

$$d = \sqrt{x^2 + y^2 + z^2}$$

With x, y and z as the position in x, y and z direction and d as the resultant distance. The time windows were again be parsed by n-back difficulty and classified using LDA and the 1 hidden layer, 100 hidden nodes ANN using 5-fold cross-validation. Finally, the averaged windows were presented for all difficulty levels to determine the effect of increased workload on to the kinematic activity.

Figure 25 shows the results for the eleven different features for ANN and LDA. It can be seen that it is possible to estimate workload directly from the controller and headset movements during the n-back task. Overall, ANN produces better results than LDA which indicates that the classification of movement windows benefits from a non-linear model. The difference for individual movements amounts to 15% for controller and 10% for headset movements. For combinations the difference for headset movements stays at 9.6% and shrinks for controller movements and a combination of all kinematic activity to 5%. A similar trend is observed for controller movement classification with 59% - 64% for ANN which achieved better classification rates than headset movements with 46.6% - 52.7%. These results appear logical since subjects used their hands to complete the task. Finally, it can be seen that combinations of all kinematic activity with 73.5% for controller and 72.5% for headset movements yield better results than the individual movements. The best results could be achieved with ANN by combining headset and controller movements with a classification rate of 82.1%.

The movement patterns averaged over all subject and all time windows for each difficulty level are displayed in Figure 26 for controller movements (left) and headset movements (right). All movement windows are corrected so that the movement starts at 0 using arbitrary units. The figure of the controller movements shows the 25% and 75% percentile in addition to the average movement as shaded error bars. Since the head traveled only very short distances and the variance was high, this would not have contributed additional information. It can be seen that the 2-back (green)

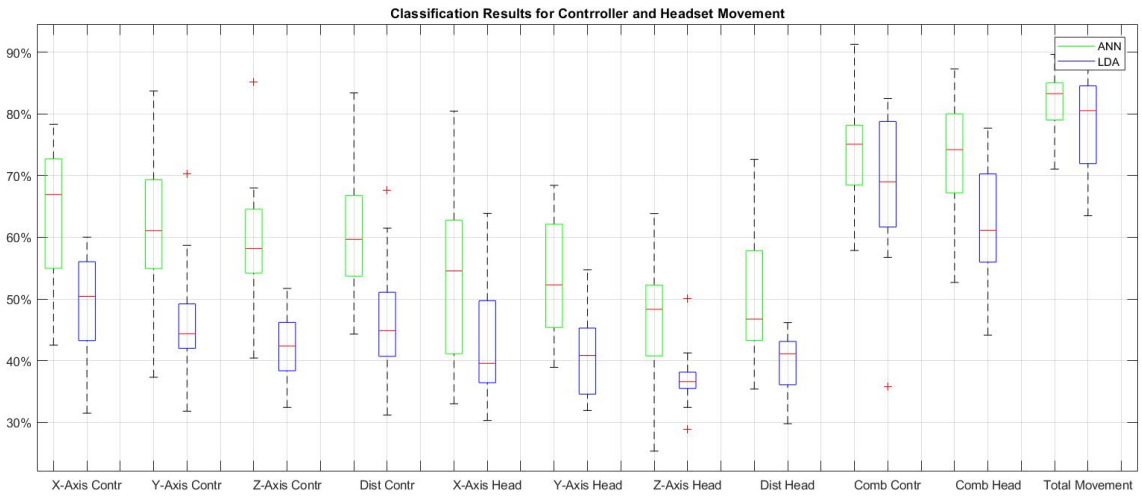

FIG. 25: Classification results for ANN (green) and LDA (blue) for movement features.

movements generally took more time than the other two conditions. The reaction time for 0-back and 1-back are nearly identical with 0-back being slightly faster. A higher peak also shows a high quantity of similar movements. Since the controller movements had a fixed travel distance, from the podium where the ball spawns to the respective receptacle, the lower peak for 2-back controller movements indicates a variation of fast and slow movements that was greater than for 1-back and 0-back. Furthermore, because headset movements did not have to travel for the n-back task, the reduced peak for 0-back shows that targets moved their heads less than for 1-back and 2-back.

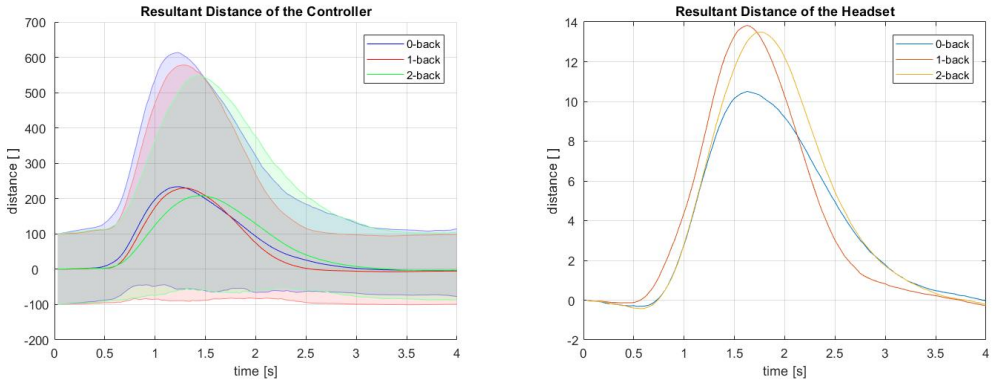

FIG. 26: Average resultant distance for controller and headset movement for each difficulty.

Figure 27 shows the stereotyped headset and controller movement paths for all subjects. The brightness of the color represents a higher chance for the movement to pass through this area. The headset movements can be seen at the bottom of each figure and the widespread controller movements are in the middle. While the movement paths are markedly different for each participant, the respective left and right paths are highly consistent for each participant. In has been demonstrated

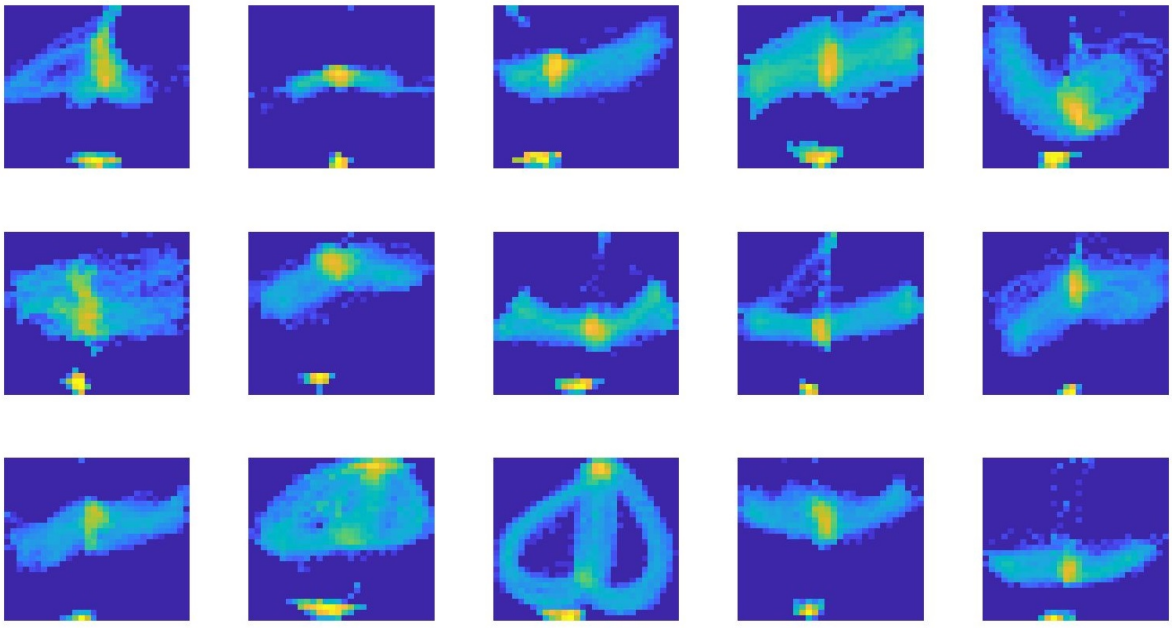

FIG. 27: Stereotypical movement paths viewed from above for all subjects.

that kinematic activity is a reliable predictor of workload and a positive correlation between movements and EEG has been proven, therefore finding a way to reduce the impact of kinematic activity on EEG is necessary. The classical high pass filter significantly reduces the correlation between contaminated EEG and task level, but also reduces the amount of data that can be used for the classification. Therefore, a new technique is presented that is named the "Warp Correlation Filter" (WCF).

6.1 WARP CORRELATION FILTER

The warp correlation filter is inspired by decontamination methods in EEG research during treadmill walking [49]. Gait artifacts in EEG during treadmill studies produce stereotypical artifacts that vary only slightly in time. Figure 26 and Figure 27 have shown that the movement windows in the VR study also depict a typical movement form that varies only slightly for increased difficulty. In order to reduce this variance a time warping algorithm is used to align the start, end and either maximum or minimum (depending which movement axis is used and in which receptacle the ball was placed in) in time. Time warping is applied to the time vector of a movement window and can either stretch or compress the time so that the already mentioned time points match. Figure 28 shows an example on how time warping works.

The top figure shows the original and time warped resultant distance of a controller during one movement. The bottom figure shows original and warped time vector of said controller movement. Time warping was applied so that the start, end and maximum of that movement is on a predefined position. The slope of the respective time vector is bigger before the predefined position, the time is running faster, so that the curve aligns with the specifications. Since the same amount of time has to pass in the warped time vector vector, just as it was in the non-time warped version, the time has to move slower after the alignment in order to compensate for it.

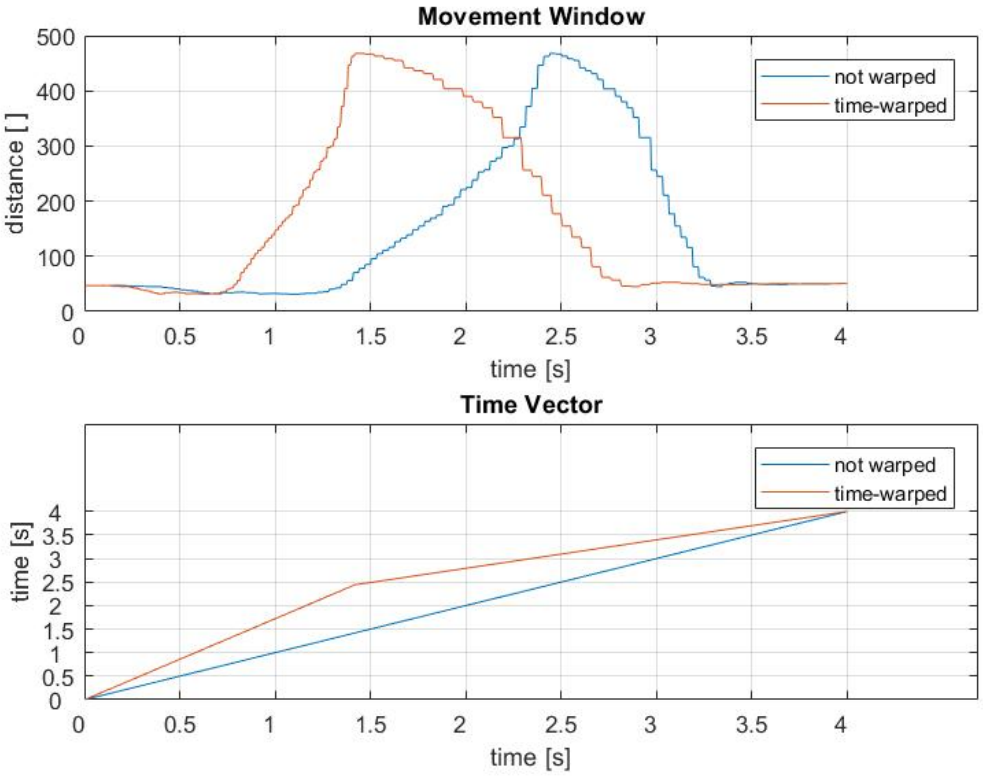

FIG. 28: Example for time warping for a time vector (top) and a movement window (bottom).

When all movements are time warped as shown in Figure 29, and align to each other, the time vector of each movement window is applied to the respective EEG channels so that they are also time warped. In order to decontaminate each EEG window, the corresponding movement window was correlated to all other movement windows. The EEG data corresponding to the highest correlated movement windows was averaged and subtracted from the original warped EEG window. Then the warped EEG was dewarped into its original condition.

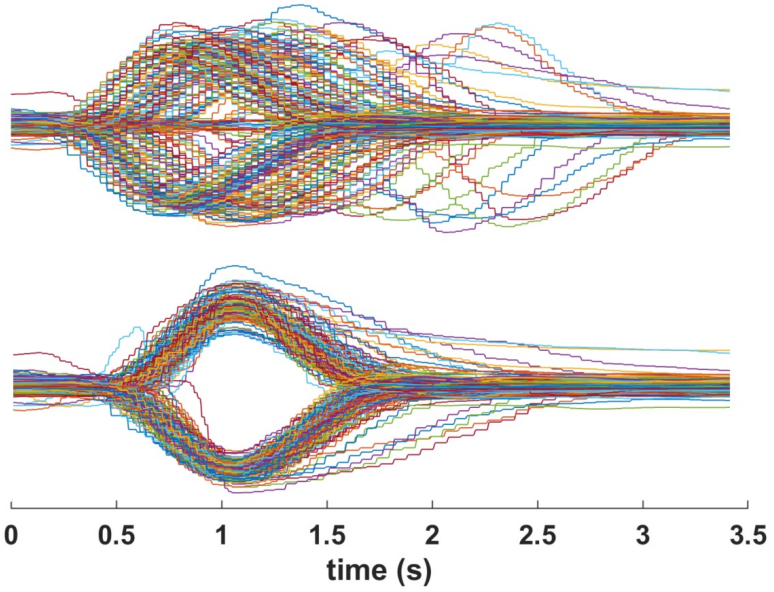

FIG. 29: Left/right controller movement for one example subject during an experiment before and after time warping.

The algorithm has two parameters that can be modulated. The first parameter is the number of highest correlating windows that are used to decontaminate EEG. Using too many windows would result in windows that are not similar enough to the contaminated window and therefore do not reduce any noise or add additional contamination. Selecting too few windows would result in contamination of individual movement uncorrelated artifacts that could have been averaged out with sufficient other channels. The second parameter is the type of source movement data the algorithm uses the decontaminate EEG. The higher the similarity between the source and EEG the better for the denoising. Figure 30 shows a similar plot that has been portrayed in Chapter 4 comparing the Pearson correlation coefficient between the raw data and the decontaminated signals. The WCF algorithm used the resultant controller distance as source movement and subtracted the 6th highest correlation windows. It can be seen that the correlation reduction drops significantly ($p<0.001$) for all channels with WCF.

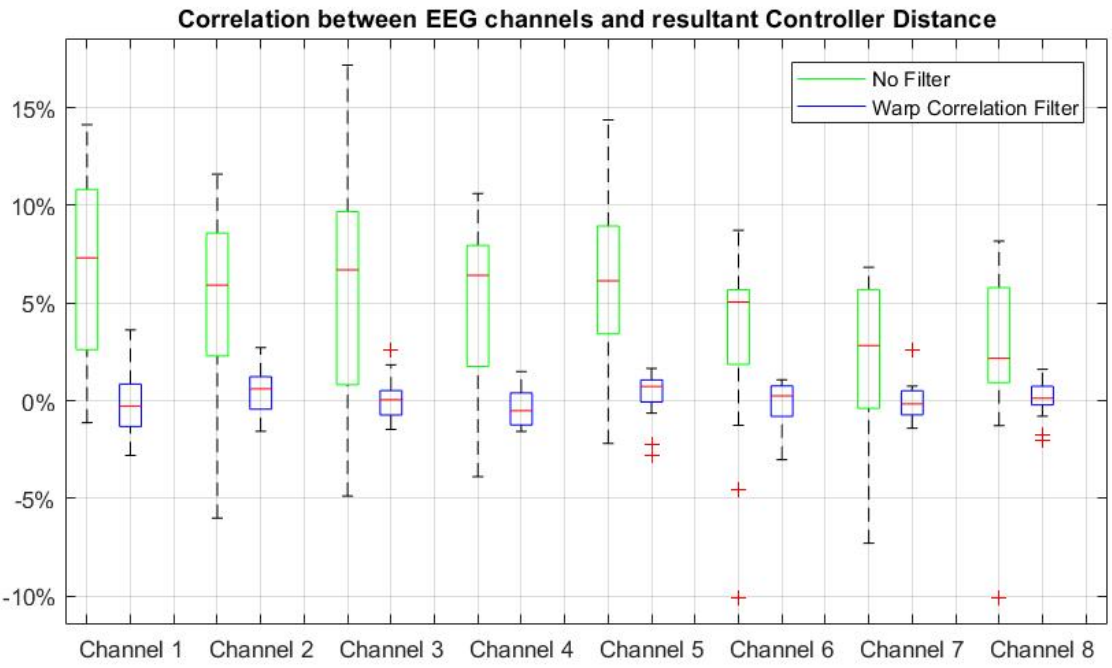

FIG. 30: Pearson correlation coefficient between EEG channels and resultant distance with WCF (blue) and without WCF (green).

Figure 31 shows the different variation of parameters. On the right side the raw signal has been compared to using the top 4 to 20 highest correlating windows for decontamination. On the left side the raw signal has been compared to using different movement sources for decontamination. Since headset movements did produce worse results only the z-axis (right/left movements) and the resultant distance has been tested. It can be seen that band power is reduced significantly. In order to actually show the difference between the parameters, the band power level of the raw data has been cut off at 5 $\frac{Volt^2}{Hz}$ but would go to 55.4 $\frac{Volt^2}{Hz}$. Comparing the use of the highest correlating windows reveals that only minor differences between the parameters. The best results could be achieved by using the six highest correlating windows with an averaged band power level of 0.97 $\frac{Volt^2}{Hz}$. The movement source evaluation depicted a greater difference between the sources with the hand movement generally having a lower band power level. The lowest average band power could be achieved using the

resultant controller distance with 0.97 $\frac{Volt^2}{Hz}$.

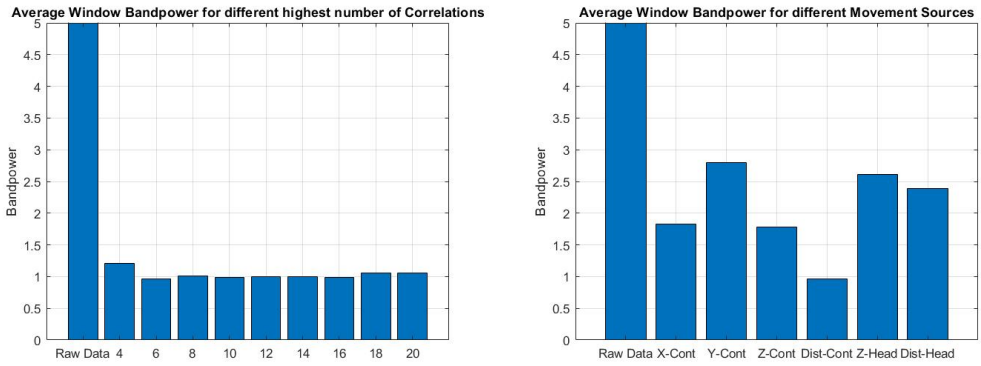

FIG. 31: Parameter testing for WCF with different number of highest correlating windows (left) and different movement sources (right).

It is important to note that the WCF can be applied multiple times in order to decontaminate artifacts that are based on different movement sources and/or directions. Figure 32 shows the band power levels averaged over all subjects when using WCF with the resultant controller distance and with another application using the remaining movement sources, controller movement in x, y, z axis and headset movement in z-axis, as well as resultant distance. While the decrease in band power for another application of WCF is considerably reduced, in this case because headset and controller shared similar movement patterns during the experiment, it is still possible to further reduce the band power. Using WCF a second time with the controller movement in z-direction further improves the average band power to $0.75\frac{Volt^2}{Hz}$. It can also be seen that using the controller movement in x-direction or using the headset movements increases the band power level and therefore adds contamination to the EEG signals. Since the reduction for using WCF multiple times is considerably less and to avoid increasing the computation time of the algorithm, the filter will only be used once for the forthcoming classification scenarios.

Using this new algorithm for movement artifacts decontamination allows the use of the the full frequency range from 0 - 30 Hz in the workload estimation problem.

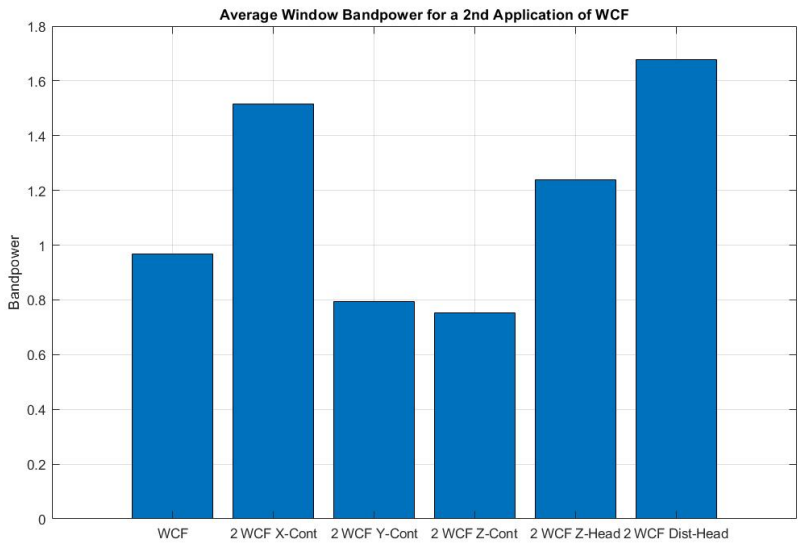

FIG. 32: Second application of WCF using the remaining movement sources.

This is accomplished by adding the sum of the delta band (0 - 4 Hz) to the previously selected features and using WCF right after the the bipolar derivation. Figure 33 shows the classification results for LDA with the new setup and otherwise the same parameters as before. It can be seen that the adjustment had a generally positive impact on performance. The average classification result for three classes could be improved by 2%. Similarly to the previous results, classes 1 and 3 preformed better than class 2. The results for the two classes generally improved by a similar amount by 1.5% for 0-back vs 2-back, 0.2% for 1-back vs 2-back and 0.8% for 0-back vs 1-back.

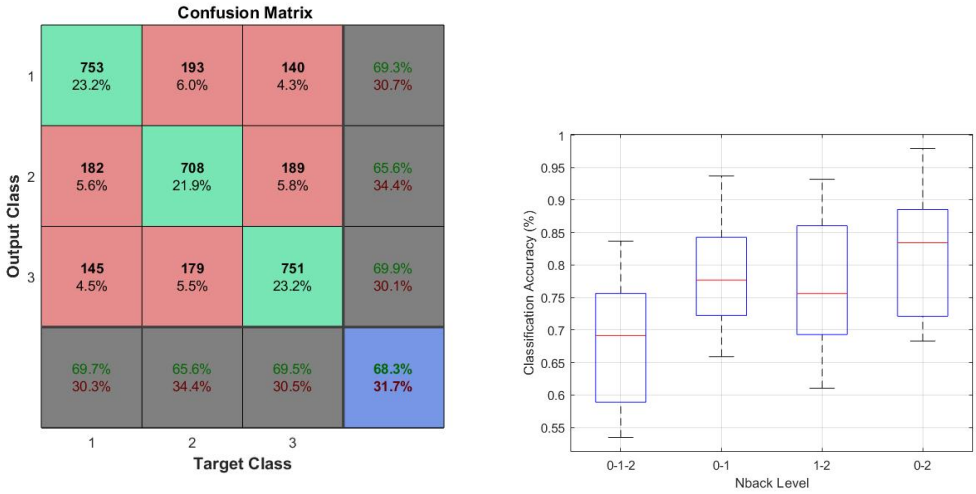

FIG. 33: Confusion matrix for three classes and boxplot for three and two classes with LDA classifier and WCF.

The results for ANN are displayed in Figure 34 and also show a slight increase for three classes of 2.2%. The two class classification increased by 1.7% for 0-back vs 2-back, however decreased slightly by 0.9% for 1-back vs 2-back and 1.5% for 0-back vs 1-back.

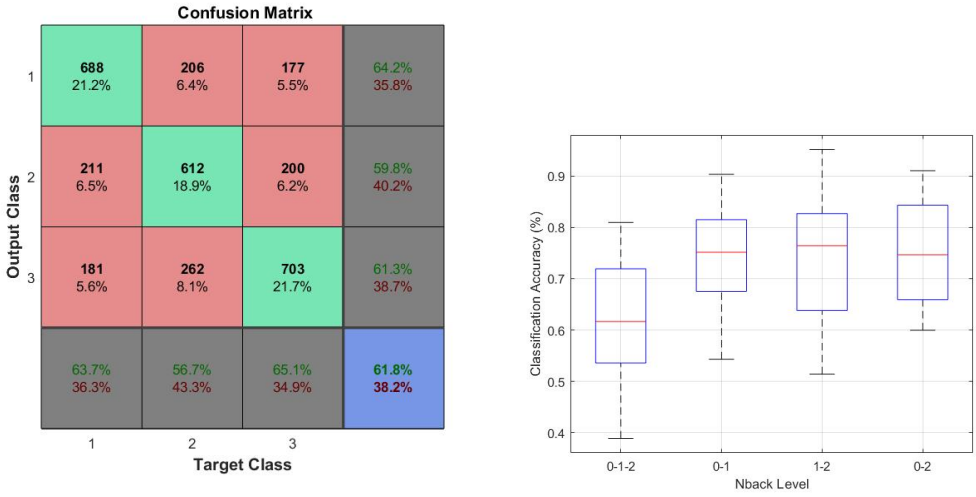

FIG. 34: Confusion matrix for three classes and boxplot for three and two classes with ANN classifier and WCF.

6.2 DISCUSSION

It has been shown that it is possible to estimate workload using kinematic information during the n-back task. The classification using ANN performed better than LDA for all movement features in workload estimation. Individual controller movements performed generally better (58% - 68%) than headset movements (48% - 55%). Combing individual movements to groups could greatly increase the results for both controller (77%) and headset (75%) movements and also reduce the difference between both groups. Combing both groups resulted in a classification accuracy of 82.1% and could outperform EEG by more than 15%. This indicates that there is more information encoded in the controller movements than just the reaction time. Particularly, subjects subconsciously performed different head movements when subjected to higher task difficulty, despite that headset movements were not involved in the n-back task.

The combination of multiple features increases the amount of information about how the subject is moving for increased workload. The increased variance of the controller movements, that is shown in Figure 26, is one example of how changes in workload affect behavioral output. With this information and the high correlation between controller movement and EEG shown in Chapter 4 it would not be possible to claim the estimation of workload purely by EEG without a denoising algorithm to reduce the influence of movement contamination.

Inspired by treadmill studies, a new filter, called Warp Correlation Filter, was developed that used the controller movement tracking to decontaminate EEG. Two parameters of WCF were tested in their ability to reduce the contamination and the best ones were selected. While it is is possible to use the filter multiple times with different parameters the noise reduction for additional uses is significantly reduced. Therefore WCF was only used once on the EEG features. The filter was able to

reduce the correlation between controller movement and EEG significantly, in contrast to a highpass filter making it possible to use the delta band (0 - 4 Hz) as features for the workload classification. Utilizing the filter and enabling the delta band as features in the EEG based workload estimation increased the result slightly by 2% for LDA and 2.2% for ANN. Despite reducing movement information that was positively correlated with the task, the WCF led to performance gains due to the inclusion of the uncontaminated delta-band information.

CHAPTER 7

ESTIMATING COGNITIVE WORKLOAD USING EMG ACTIVITY

In several studies [39–41] it has been has been shown that muscle tension is modulated by workload and produces EMG signals that can significantly contaminate EEG across all frequencies, predominantly starting around 20 Hz. Therefore, it should be possible to use EMG activity for workload estimation. Since the practical bandwidth of scalp EEG is approximately between 60 and 80 Hz, compared to the roughly 300 Hz bandwidth of EMG, it is reasonable to presume that any task-related activity above 60 Hz is predominantly due to EMG activity.

To be confident that residual EEG was not present, the EMG surrogate used for classification was the frequency band from 90-118 Hz. The frequency range of 118 - 128 Hz was excluded due to the harmonic, a positive integer multiple, of the power line hum artifact at 60 Hz that could interfere with the classification. Since EMG is not characterized into traditional frequency bands as with EEG, three different EMG bandwidths were tested ranging from from 5 Hz to 28 Hz resolution. Specifically, features consisted of four 5 Hz bands and one 8 Hz band, two 10 Hz bands and one 8 Hz band and one 28 Hz band from 90 - 118 Hz. Since it was very likely that longer windows would produce better results based on previous evaluations, the window size for features will again be 4 seconds. The bipolar derivation as well as PCA, ICA, Hampel's outlier filter and the parameters for Welch's method, were consistent with the previous EEG analysis. The features will again be classified using LDA and ANN with a 5-fold cross validation.

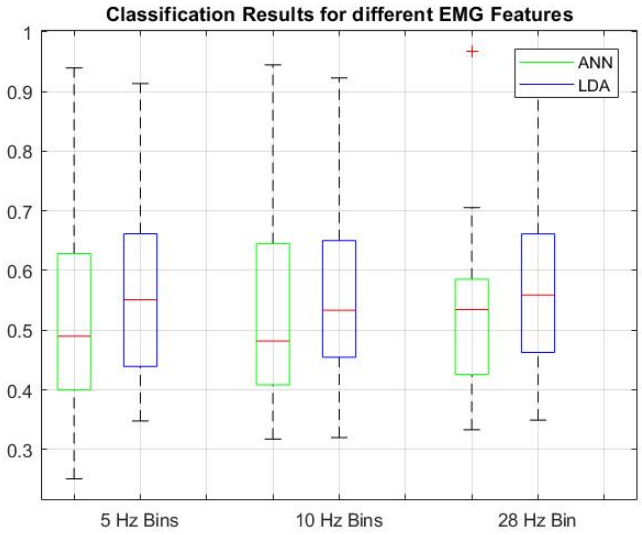

FIG. 35: EMG feature testing.

Figure 35 shows the result of the selected features. It can be seen that all features perform similarily in a range between 50 - 53.4% for ANN and between 56 - 57% for LDA. ANN features perform slightly better with 53.5% classification accuarcy using only one 28 Hz feature. The best results could be achieved using the 28 Hz bin from 90 - 118 Hz with 57% using the LDA classifier. This suggests that EMG can be used to classify workload and while EMG does not show the same predictive power as EEG, it is still far above chance level. Since EMG and EEG overlap in their frequency ranges it is unclear which signal has been used as a predictor in the frequency range of 0 - 30 Hz. In order to ensure that only EEG is used, another technique is needed to decontaminate EEG of EMG signals. Similar to WCF, the algorithm must cancel the EMG signals in the frequency range of EEG using a source EMG signal. Since the distribution of EMG signals over all frequency bands is linear [78] a linear regression model can be used.

The regression model estimates the correlation between the 90 - 118 Hz range and the 12 - 58 Hz range and then subtracts the estimate from the EEG signal. The lower limit of 12 Hz was chosen because a lesser value would interfere with the positive

correlation between the theta band and workload. Figure 36 (left) shows a correlation analysis between the raw EEG data and the difficulty level of the n-back task. The curve was averaged over all subjects and all EEG channels. It can be seen that there is clearly a positive relationship between EEG and workload between 4 and 8 Hz. Then the correlation drops to negative values until 20 Hz and then increases again to positive values for bands higher than 28 Hz. Using linear regression correction for values under 12 Hz (the most negative correlation coefficient) would cancel out the positive correlation peak between 0 and 12 Hz. The figure on the right shows a similar graph which details the correlation between the lower frequency 0 - 58 Hz EEG features and the higher frequency 90 - 118 Hz EMG features. The curve was averaged over the EMG frequency bands in that range, the subjects and the channels. It can be seen that the curve follows a similar pattern with a drop in correlation at 10-11 Hz. But compared to the first graph, the correlation values are a lot higher and the differences between theta and alpha bands are a lot smaller, which indicates EMG contamination.

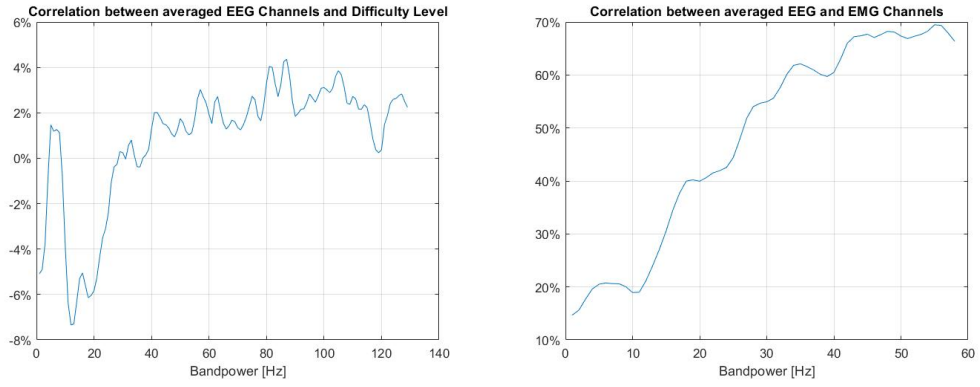

FIG. 36: Correlation between EEG spectral power frequency levels and the n-back difficulty levels (Left) and EMG spectral power levels (Right).

7.1 LINEAR REGRESSION DECONTAMINATION

Linear regression analysis is the study of linear, additive relationships between variables. The dependent outcome variable Y is being predicted by multiple independent predictor variables $X_1, ..., X_\mathrm{k}$. With the constants $b_1, ..., b_\mathrm{k}$ the equation for Y denotes to:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k$$

In this case Y is the predicted EMG proportion in EEG and $X_{90}, ..., X_{118}$ are the EMG bands from 90 - 118 Hz. The equation for the corrected low frequency EEG components $\hat{X}$ with $i = 12, .., 58$ as the low frequency bands denotes then to:

$$\hat{X}_i = X_i - (b_{0,i} + b_{1,i} X_{90} + b_{2,i} X_{91} + ... + b_{29,i} X_{118})$$

This algorithm was applied on each frequency band between 12 and 58 Hz using the same dataset and correlation analysis as before. This technique nullified the correlation between the selected EEG bands and EMG as Figure 37 (right) shows.

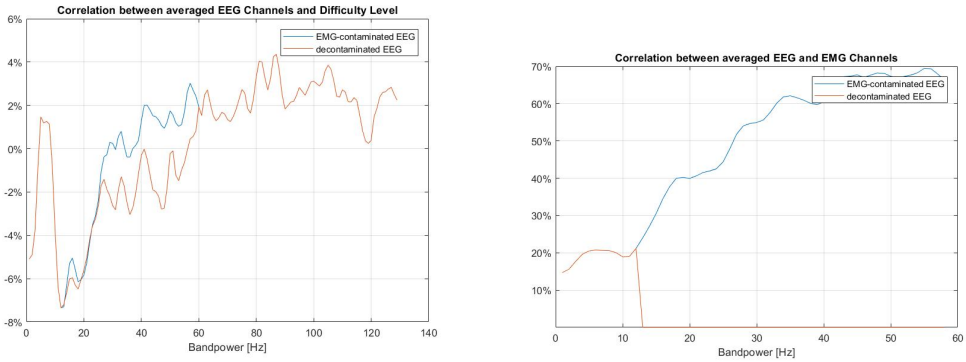

FIG. 37: Correlation between regression decontaminated EEG spectral power frequency and the n-back difficulty Levels (Left) and EMG spectral power levels (Right).

The correlation value of decontaminated EEG drops and stays at 0% after the 12 Hz band. The correlation between the decontaminated EEG and the difficulty level

(left) reveals the influence of EMG at frequencies between 26 Hz and 56 Hz. In the original dataset the positive correlation of EMG to the difficulty level overwrote the negative correlation of EEG. The decontaminated EEG shows now a slightly negative correlation in that frequency range and a more negative correlation in the range of 14 Hz to 20 Hz.

Using this algorithm allowed the addition of the complete frequency range from 0 - 58 Hz as features for the EEG-based workload estimation. The frequency bands around 60 Hz have again been excluded due to the power line hum artifact. The sum of the gamma band (30 - 58 Hz) was added to the previously selected features and the regression decontamination was performed after using Welch's method. The classification results are shown in Figure 38 for LDA with the new setup, maintaining the aforementioned processing parameters.

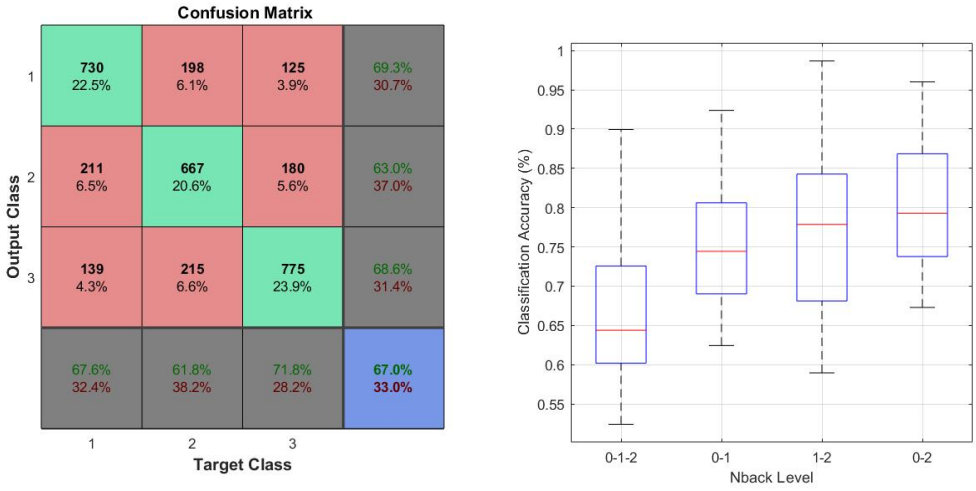

FIG. 38: Confusion matrix for three classes and boxplot for three and two classes with LDA classifier, WCF and regression denoising.

It can be seen that the results dropped into comparison with just using WCF but performed similar to classification without any extra decontamination technique. Compared to EEG without decontamination, the linear classification achieved a results of 67%, which is a slight increase by 0.7% for three class classification. The

classification accuracy for two classes increased by 2% for 0-back to 2-back and by 0.9% for 0-back to 1-back, in contrast the accuracy decreased decreased by 1.9% for 1-back to 2-back.

The results for the neural network achieved are shown in Figure 39. For the three class problem the accuracy was 59.3% and dropped by 0.3%. For two classes the results decreased by 0.5% for 0-back to 2-back, by 1.2% for 0-back to 1-back and by 1.9% for 1-back to 2-back.

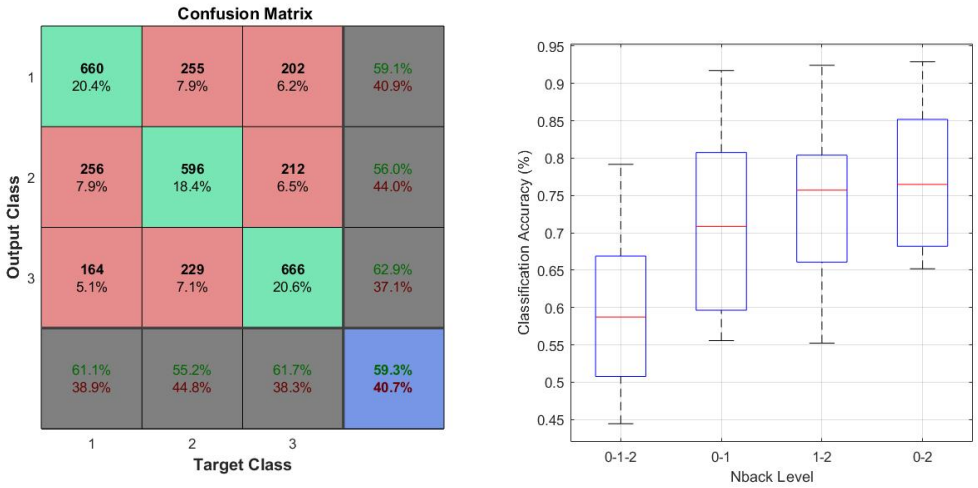

FIG. 39: Confusion matrix for three classes and boxplot for three and two classes with ANN classifier, WCF and regression denoising.

7.2 DISCUSSION

It has been shown that it is possible to estimate workload using EMG activity during the n-back task. Since there are are no traditional EMG frequency bands like for EEG, three different band sizes (5 Hz, 10 Hz and 28 Hz) have been tested at a frequency range from 90 - 118 Hz. This range was picked to ensure no EEG signals would be contained as frequency components of EEG signals are limited to approximately 60 Hz. EMG as an estimator of workload performed worse than EEG or movement features with an average result of 57% for LDA.

With this information, the overlapping frequency ranges and the high correlation between EMG and EEG features make it impossible to assess the contribution without incorporating an EMG denoising algorithm. Assuming a linear relationship between EMG activity in different frequency bands, a linear regression model was developed that estimated the low frequency (12 - 58 Hz) EEG components from high frequency EMG surrogate signals (90 - 118 Hz). Since the relative band power of EMG is decreased below 20 Hz, and in order to not eliminate the positive correlation between the theta band and workload, the contamination algorithm was not applied on the complete frequency range of EEG. The lower limit of 12 Hz was selected since it was the point of lowest correlation coefficient between EEG and EMG. The regression-modeled output was subtracted from EEG in order to cancel the correlation of EEG and EMG. This enabled the application of decontaminated gamma to the classification algorithm.

Utilizing the regression decontamination and enabling the gamma band as features in the EEG based workload estimation decreased the result slightly by 0.7% for LDA and 0.3% for ANN, compared to EEG without any decontamination algorithm. The correlation analysis showed a rise in correlation between EMG and EEG beginning at 12 Hz compared to the 20 Hz that is suggested by literature [42]. This indicates EMG contamination can begin in the beta band, and since EMG is positively correlated with workload it would support a linear classifier. As regression denoising removes the EMG influence in this band, it also removes the positive correlation. Thus, explaining a slight drop in classification accuracy compared to WCF.

CHAPTER 8

EEG SIGNAL ANALYSIS

The previous chapters analyzed two sources of signals that contaminate EEG during workload studies while the contamination itself was modulated by workload. This contamination is another reason why numerous workload studies do not establish a consistent frequency range and the behavior of EEG in workload classification, aside from including the theta and alpha band. This chapter will focus on the influence of workload on EEG in the frequency range of 0-58 Hz by analyzing the features that have been used for the classification. The first analysis will use the standard score (or z-score) to normalize the EEG features in order to see the change in time and for each difficulty level. The standard score z of the raw score x with the mean μ and the standard deviation σ denotes to:

$$z = \frac{x - \mu}{\sigma}$$

The standard score has been calculated for two feature sets. The first set of features was created by using two notch filter between 58 - 62 Hz and 118 - 122 HZ and then transforming the 4 second EEG windows into frequency components using Welch's method with 256 FFT points, 128 window size and 60% overlap. The second dataset uses the same Notch filters, bipolar derivation, WCF and regression decontamination, while utilizing the same parameters for Welch's method. The standard score was calculated by computing the mean and standard deviation of the power values for each subject, each channel and each frequency band, and subtracting it from the average frequency power value for each subject, channel and frequency band, respectively. Figure 40 shows the standard score using only the notch filter averaged

over the delta, theta, alpha, beta, gamma and EMG power band and averaged for each trial and each subject.

Each figure shows the standard score for all of the eight channels as it evolves during the average trial. Each four second mark represents one stimulus presentation. It can be seen that the standard score for first two (0 - 8 seconds) stimulus presentations is considerably elevated at the beginning of the trial and decays over time. This pulse-decay effect is also visible for the next couple of stimuli as it slowly subsides. The standard scores of the first two windows increased with higher frequency bands. While in delta, theta and alpha band the average elevation is between 0.43 - 0.44, beta, gamma and EMG band show an elevation of 0.57 - 0.7 on average. Also, it takes the gamma and EMG band nearly 30 seconds to decrease to a steady baseline and the delta and alpha band only 20 seconds. The pulse-decay effect's behavior therefore resembles that of an exponential decay function.

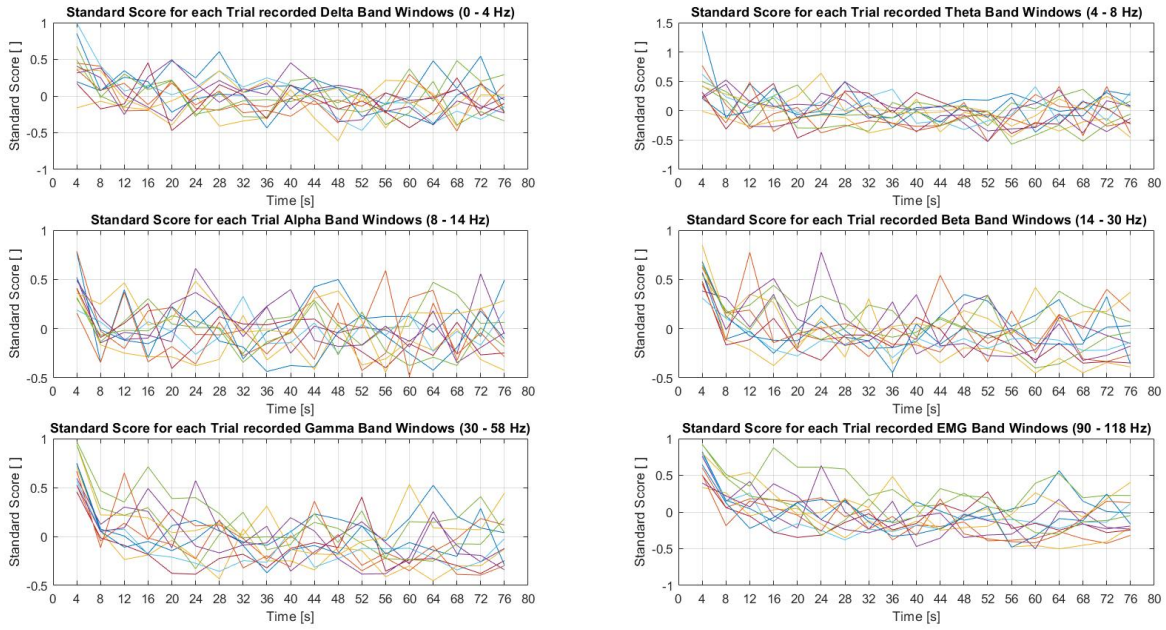

FIG. 40: Standard scores for raw data averaged over each trial and subject. Each trace represents a single EEG channel.

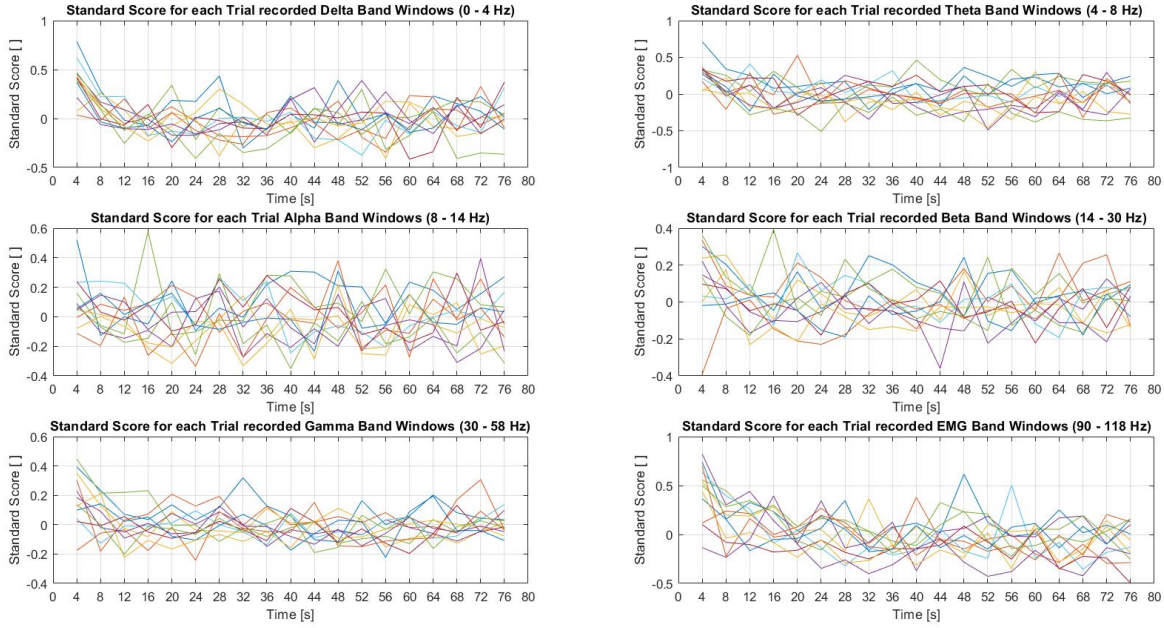

FIG. 41: Standard scores for processed data averaged over each trial and subject. Each trace represents a single EEG channel.

Comparing these results to Figure 41 reveals the effects of decontamination techniques that were used on the data. Nearly all frequency bands show a reduced pulse-decay effect in amplitude and time. The effect is still visible in the delta and theta range compared to the alpha and beta after decontamination. Since regression decontamination was not applied to the delta and theta range and, since the effect is generally stronger in higher frequency bands, it can be assumed that this effect is related to the muscle groups that produce the EMG signals. This pulse-decay effect also explains why the classification results would improve when the first windows were excluded from the feature generation. The high power values that were recorded at the beginning of each trial independently of the task difficulty level adversely impacted the classification. For the further analysis this effect will be called effect 1.

The second analysis compared the progression of the standard score over the course of the whole experiment. Both setups that are presented are identical to the

ones before. Figure 42 shows the standard score of the raw signal for each channel and effect 1 can already be identified as the standard score rises at each trial start, but additionally another transient pulse-decay effect can be seen, which increases the standard score for the complete first and second trial. The effect starts at a standard score of higher than 1 for trial 1 and subsides within the 2nd trial. It is prevalent on all frequency bands and has always a similar form. Comparing these raw EEG plots to the decontaminated ones in Figure 43 reveals that, like the average trial in Figure 41, the influence of effect 1 is considerably reduced. Only the EMG, and to a lesser extent the delta and theta band, display the stereotypical peaks at the begin of each trial. It can also be seen that the secondary pulse-decay effect is still prevalent in all frequency bands and is only slightly reduced.

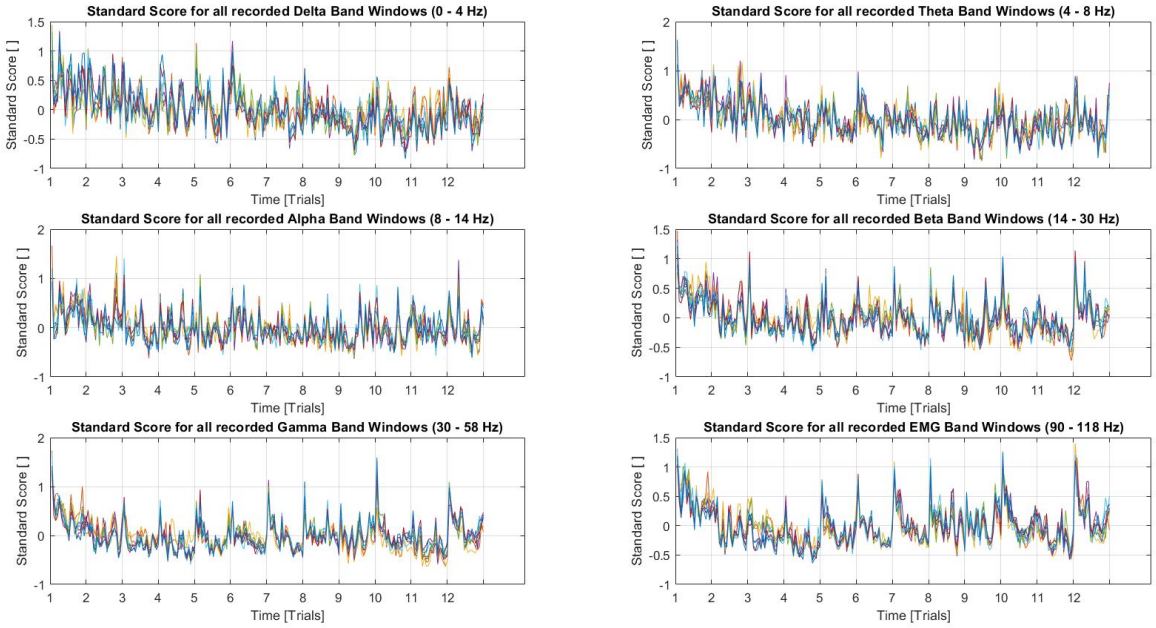

FIG. 42: Standard scores for raw data for the whole Experiment averaged over each subject. Each trace represents a single EEG channel.

The behavior of the second effect resembles again a decaying trend but the gradient is substantially smaller than for effect 1. Since the decontamination algorithms did

not affect this pulse-decay and the effect is also not influenced by the frequency, this suggests that effect 2 is not a product of movement or EMG artifact contamination.

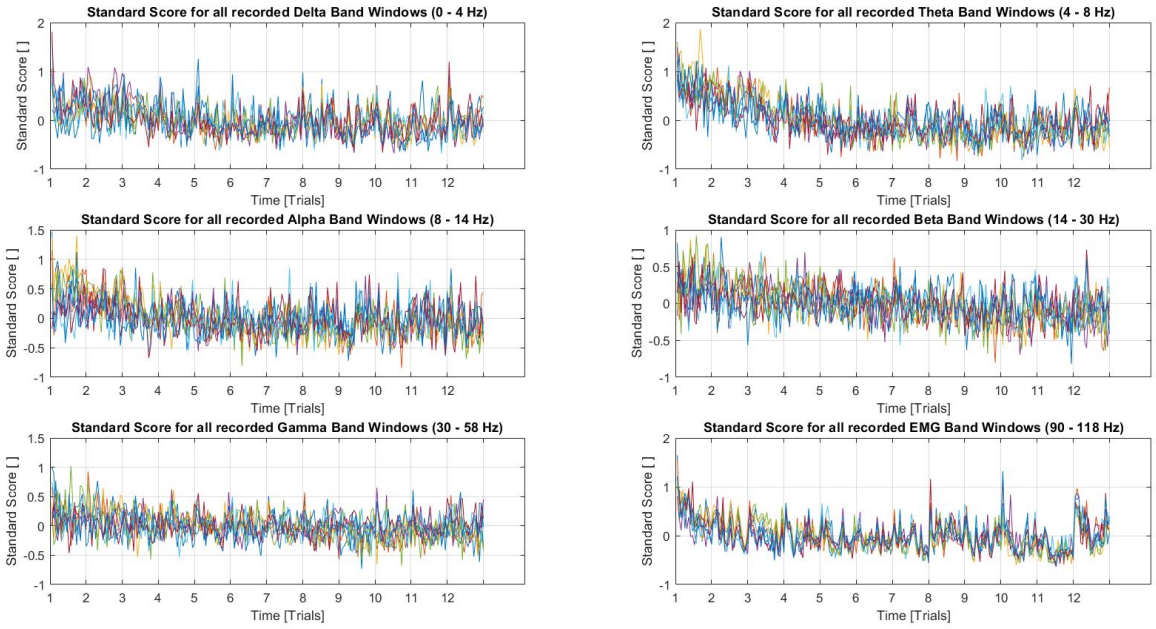

FIG. 43: Standard scores for processed data for the whole experiment averaged over each subject. Each trace represents a single EEG channel.

Since both effects are prevalent in all frequency bands it should be possible to see them in the movement features. Therefore, the variance for each four second window was calculated comprised of combined controller movements features used in workload classification. The standard score of the variance was calculated for each subject. For Figure 44 (left) the standard score was averaged over all subjects and trials and then compared to the EEG standard scores in the already shown frequency bands. It can be seen that a similar transient pulse-decay is present in EEG and movement standard scores. The progression for the whole experiment is displayed in Figure 44 (right) averaged over all subjects. Again it can again be seen that a similar effect exists that is elevated at the beginning of the experiment and decays to a consistent baseline. In trial two and four the movement standard scores deviate

slightly from EEG standard scores, possibly due to anomalies in the data. While this chapter has analyzed the effect of the workload experiment on EEG and EMG, it did not describe the effect workload levels on these features. This has a simple reason: while it is possible to make general statements to the effects of cognitive workload, the subject's individual response is unpredictable. This has been reported in multiple studies [33] when looking for statistically-significant features. A good demonstration is the individual correlation between EEG spectral power and difficulty level for each subject as presented in Figure 45. It demonstrates the correlation coefficients for each subject and highlighted for the average. The notable variance of the coefficients confirms the subject individual response, and while the average correlation analysis agrees with the findings of prior research [33, 44], the standard score analysis did not give any useful results. Another reason why the standard scores for each difficulty level were not presented was due to effect 2. Since the first two trials are elevated compared to the remaining trials for effect 2, this inherently increases the band power level independent of difficulty level, resulting in a misleading representation of EEG with respect to workload.

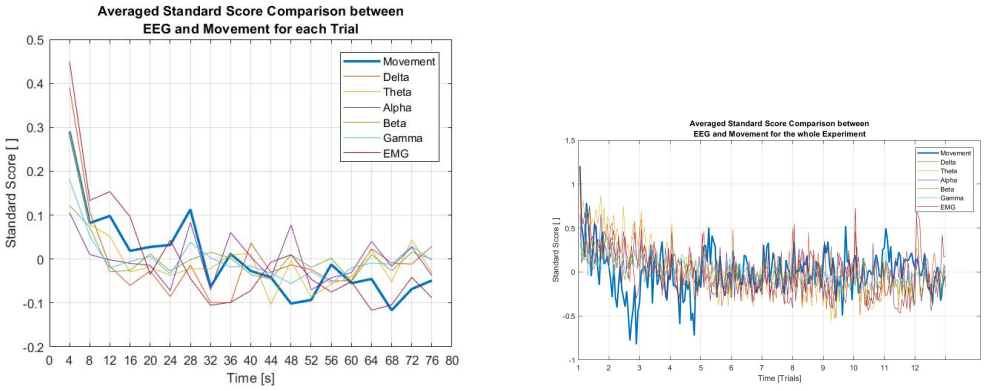

FIG. 44: Standard score for the variance of movement features for each trial (left) and for the whole experiment (right).

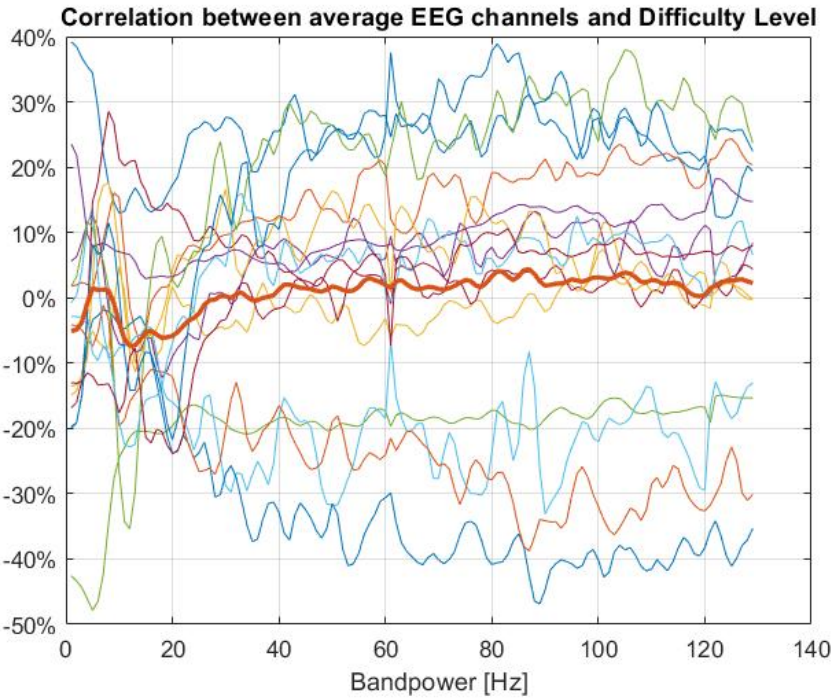

FIG. 45: Correlation between EEG band power averaged over all channels and difficulty level for each subject.

8.1 FREQUENCY BAND FEATURE ANALYSIS

In addition to the standard score feature analysis a statistical analysis was performed. To assess the most discriminable univariate features across subjects, a two-sample t-test was used to determine the percentage of participants with statistically-significant differences in spectral amplitude between the extreme workload levels of $n = 0$ and $n = 2$ for each feature. The results shown in Figure 46 were Bonferroni corrected to a significance level of $5.2 * 10^{-4}$ (0.05/(12 frequency bands x 8 channels)). Decontaminated frequency bands are indicated with the asterisks (*). Roughly 45% of the subjects exhibited a significant difference in delta activity and theta activity in channel 2 before decontamination. The most prevalent significant features after decontamination, with roughly 50% prevalence, are again in the delta and theta in channel 2, but additionally in the delta and theta band channel 8 and the beta band

channel 1. It can also be seen that the decontamination reduced the significance of gamma band features from channel 3 - 8. The plot additionally confirms the individual spread of the features for subjects as nearly all features show a significant difference for least 15% of the subjects. Finally, an across-subject classification was

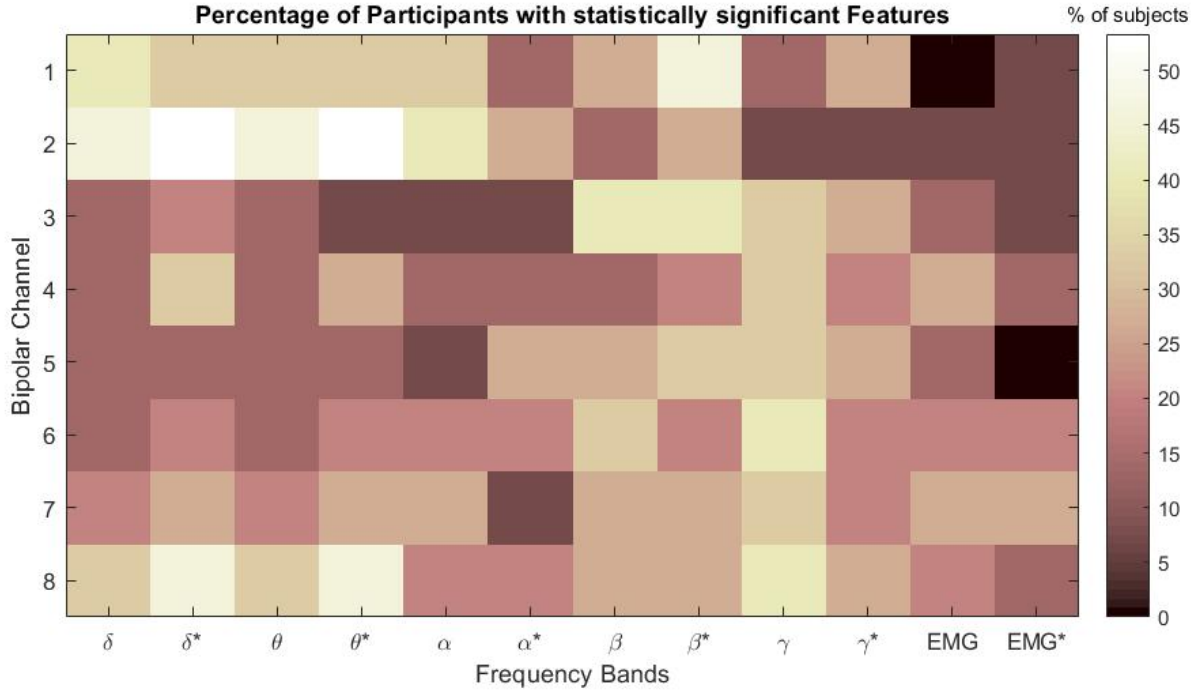

FIG. 46: Percentage of participants with statistically significant differences in spectral amplitude between 0-back and 2-back difficulty.

performed by training the classifier with the features of 14 subjects and testing it with the remaining subject. The standard parameter with bipolar derivation, ICA, PCA, regression, hampel outlier removal and WCF decontamination was chosen with ANN, LDA and 5-fold cross-validation. The results can be seen in Figure 47 on the left for LDA and on the right for ANN. Every measure is nearly exactly at chance level (33% for three class classification and 50% for two class classification). The ANN classifier failed completely for 1-back to 2-back and 0-back to 2-back condition and always performed at the chance level. The results agree with the statistical and correlation results and demonstrate the individual spread for each subject.

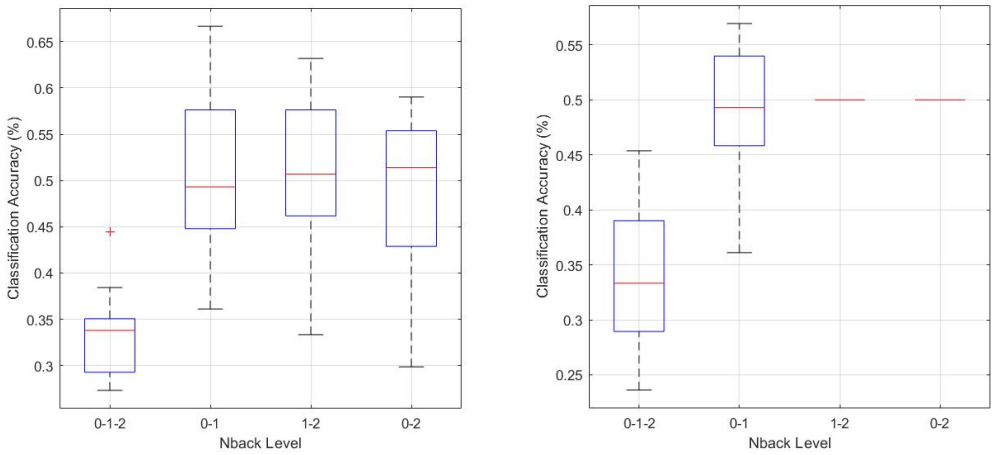

FIG. 47: Boxplot Results for across-subject for LDA (left) and ANN (right).

8.2 EEG CHANNEL ANALYSIS

While the previous section focused on statistical evaluation of the frequency bands, this section focuses on the EEG channels and the corresponding brain areas. Figure 48 presents the correlation coefficients between raw EEG and the difficulty level averaged over each subject. Each of the traces, representing one EEG channel, follows the typical pattern and displays an elevation in the theta band range and decay in the alpha band. But there are differences between the curves which depend on the position of the electrodes. Electrode 1, 2 and 3 were positioned over the frontal lobe and show the greatest correlation (up to 8%) in the theta range of all channels. Even though the correlation coefficient of the central motor cortex electrodes 4 and 5 peaks only at 2% it is still a considerable elevation compared to the alpha range at -7%. The parietal electrodes show only a slight increase and stay in the negative correlation range. In contrast to the theta range, electrode 2 and 3, positioned at each hemisphere in the frontal lobe, display the most negative correlation for frequencies higher than 20 Hz to increased workload. All other electrodes, including the central frontal electrode 1, feature a higher, positive correlation for higher frequencies.

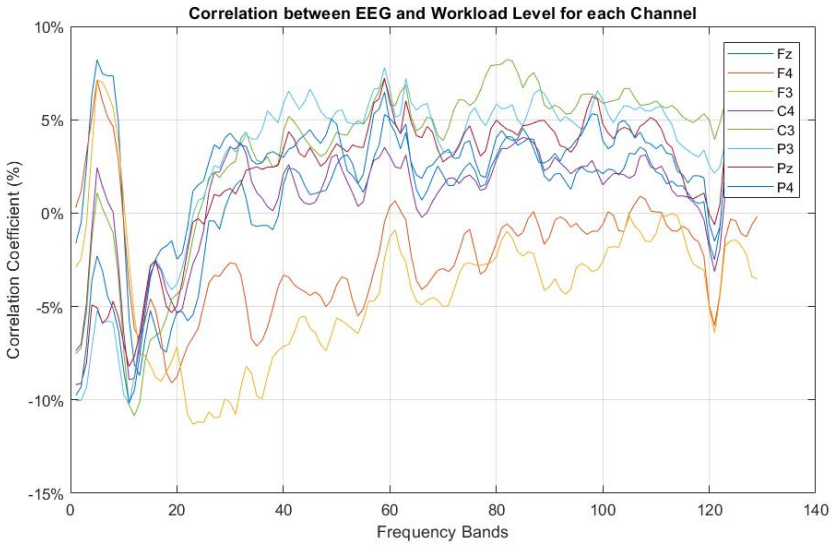

FIG. 48: Correlation coefficient for each EEG channel.

In addition to the correlation analysis, a statistical two-sample t-test analogous to the frequency band analysis was performed. In this case the aim was to highlight the significant differences between the extreme workload levels $n = 0$ and $n = 2$ for each electrode channel. In contrast to the previous analysis 1 Hz frequency bands were tested. Results were Bonferroni corrected to a significant level of $4.8 * 10^{-5}$ (0.05/(128 frequency bands x 8 channels)) and then averaged over all frequency bands. Figure 49 displays the percentage of subjects that showed a significant difference averaged over all frequency bands. It can be seen that all channels perform similarly, where about 30% of the subjects show significant differences. Channel F3 performed best with 32.2%; the worst result obtained P4 with 24.4%. It is noticeable that electrodes on the left brain hemisphere (F3, C3, P3) overall performed better than their counterparts on the right hemisphere (F4, C4, P4).

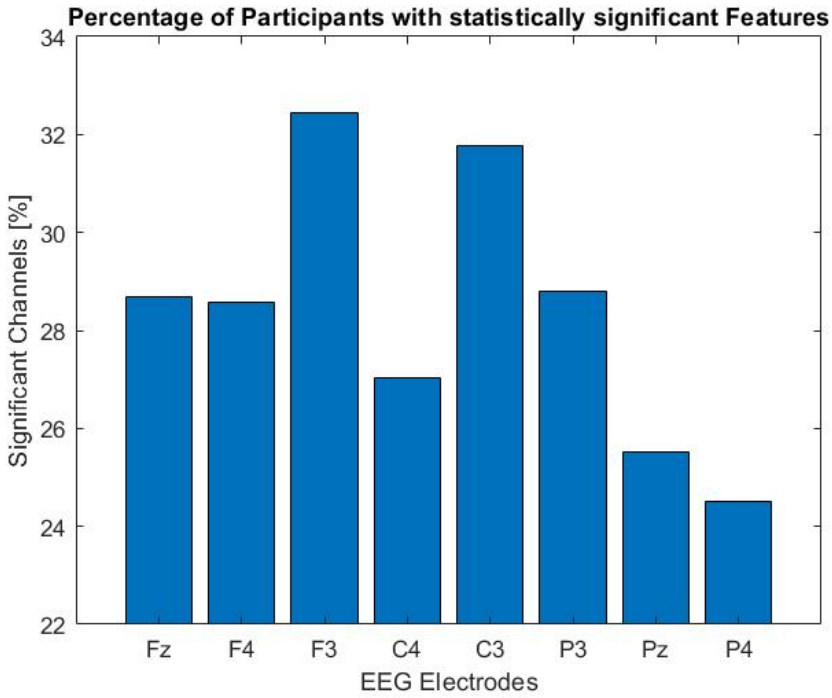

FIG. 49: Percentage of participants with statistically significant features for each electrode channel.

8.3 DISCUSSION

In this section, EEG features and the influence of the decontamination were analyzed. Standard score analysis revealed two effects that decay EEG features from the beginning of the experiment (effect 2) and at the beginning of each trial (effect 1). Effect 2 has already been reported and Berka et al. attributed the "relatively high engagement during the first level of each task and a decrease in engagement for the second level of task difficulty," to, "an initial task adaptation or novelty response" [44]. This appears to be a reasonable explanation, as most participants were neither familiar with the n-back task, nor with the VR setup. Pre-task introduction phase for this experiment was intentionally kept short to maximize the experimental portion of the session and to decrease the chance of simulator sickness due to extended time in VR. The effect also applied to the participants' arm movements which could

also be attributed to the unfamiliarity of the subjects with the experimental setup.

Effect 1 temporally overlaps the transition effects, the adaptation of the human body to changing demands. However, it does not match a reaction of the human physiology from rest to increased workload. An expected transition effect would rather start with decreased EMG activity that would increase with workload. Yes, in this instance the opposite is true. Therefore, heightened attention, a preparatory state of focus of the subjects immediately before the task, is a logical explanation. The countdown could have increased the anticipation of subjects, which produced a similar muscle tension activity as increased workload. This EMG signal could have masked the EEG across the entire frequency range, and would also explain why decontamination of EMG artifacts would reduce the effect. In addition to the standard score feature analysis a more in-depth analysis of the features for each subjects was performed. Correlation analysis, a two-samples t-test and across subject classification confirmed the large individual spread of features for each subject, which highlighted the need for subject-dependent classifiers.

Analyzing the correlation coefficient between EEG features and difficulty level revealed that electrodes positioned at the frontal lobe showed a higher peak correlation in the theta band compared to electrodes at the motor cortex and parietal lobe. The frontal lobe is associated with decision making [79] and researchers have found a connection between increased theta activity and memory load in working memory tasks [80]. For higher frequency bands electrodes placed on each hemisphere of the frontal lobe showed a negative correlation while all other electrodes show a positive correlation. Facial and eye muscle artifact contaminate EEG channels at the frontal lobe. This negative correlation could indicate that the EMG contamination in the motor cortex and parietal region was produced by a different muscle than in the frontal region. Muscle artifacts in the frontal lobe are commonly associated with facial and/or eye movements and frontalis muscle tension, while artifacts in the in

the temporal regions can be attributed to temporalis muscles [42].

The statistical analysis revealed that electrodes positioned on the left hemisphere showed increased significant difference between the extreme workload levels 0-back and 2-back. The increased significance on the left side can be explained with the representation of movements in the human brain. All subjects used their right hand to perform the n-back task; movements in the right are are processed by the motor cortex in the left hemisphere. Subjects performed different movement patterns for increased difficulty levels and therefore a more significant difference is visible on the left hemisphere.

CHAPTER 9

COMBINING FEATURES

The goal of this last chapter is to create the best performing features by combining movement, EMG and EEG features to get the best possible estimation of cognitive workload. For the first setup, EEG and EMG features were combined to create a workload estimation based on electro-physiological signals. Bipolar derivation, WCF with the resultant controller distance as source using the 6 highest correlations, PCA, ICA, regression denoising from 12 - 58 Hz and the Hampel outlier filter were applied. EEG and EMG features were decomposed to frequency components using Welch's method with 64 FFT points, 32 point windows and 60% overlap. Frequency bands were selected as follows: delta (0 - 4 Hz), theta (4 - 8 Hz), alpha (8 - 14 Hz), beta (14 - 30 Hz), gamma (30 - 58 Hz) and EMG (90 - 118Hz). All features were then combined and parsed by n-back difficulty. LDA and ANN were used as classification algorithms with 5-fold cross validation.

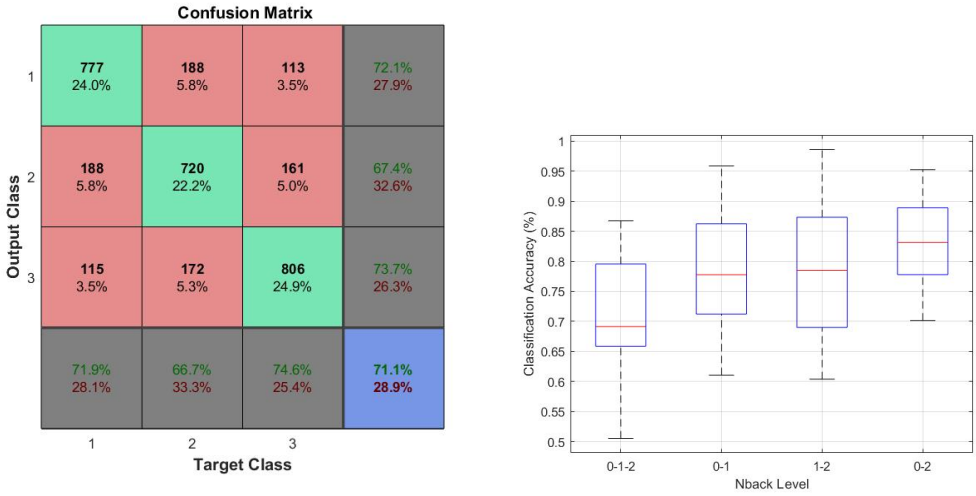

FIG. 50: Result of combining EMG and EEG features with LDA classifier.

The results of the electrophysiological combination with LDA can be seen in Figure 50. The results show an increase in classification accuracy compared to just EEG or just EMG. It has previously been shown by workload studies [6] that combining multiple physiological features increased the classification accuracy of the best estimator. The accuracy for 3 classes resulted in 71.1% and increased by 4.1% compared to EEG, and by 14.1% compared the EMG. The results for 0-back to 1-back and 1-back to 2-back were at 78.1%, while 0-back to 2-back performed best with 82.4%.

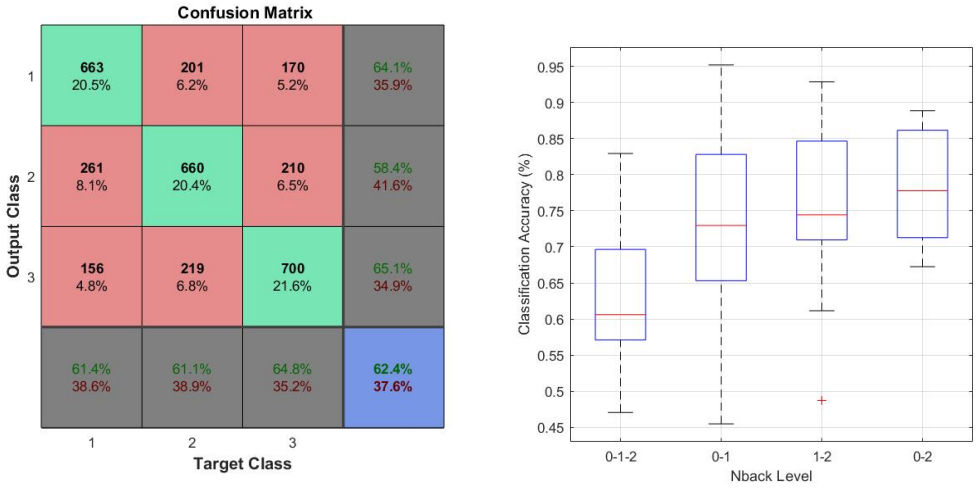

FIG. 51: Result of combining EMG and EEG features with ANN classifier.

ANN performance has been consistently lower for electrophysiological signals. While the combination of EMG and EEG using ANN could increase the results compared to the single measures, it also increased the difference between LDA and ANN to 8.7%. The classification rate for three classes resulted in 62.4% and increased by 3.1% compared to EEG only, and by 9% compared to EMG only. The results for two classes were at 73.1% for 0-back to 1-back, 76% for 1-back to 2-back and 78% for 0-back to 2-back.

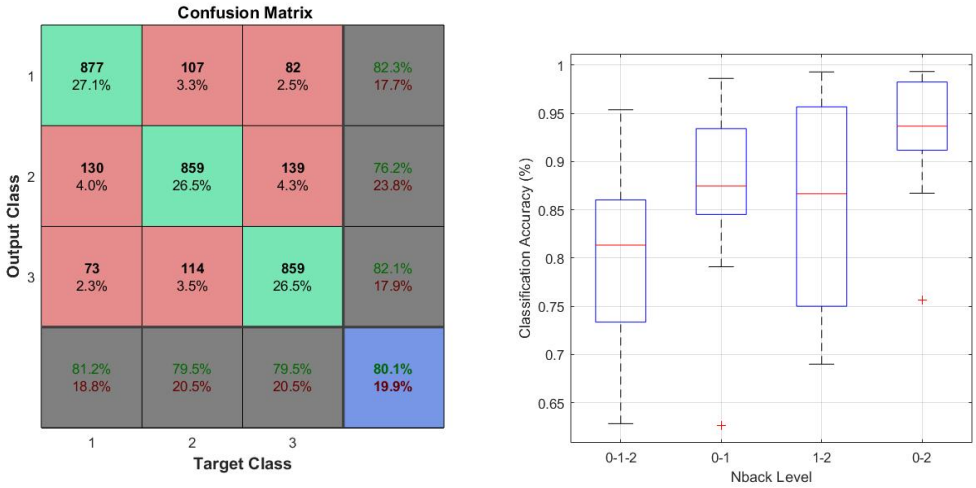

FIG. 52: Result of combining EMG and movement features with LDA classifier.

The second setup adds the Hampel-filtered movement features but uses only the EMG frequency range between 90 - 118 Hz. Combining EMG and movement features resulted in a slightly improved classification accuracy compared to just movements. Figure 52 shows the classification accuracy for LDA. For three classes it resulted at 80.1% and stayed the same compared to movement features and increased by 23.1% compared to EMG. The two class classifier could achieve 89% for 0-back to 1-back, 88.3% for 1-back to 2-back and 91.6% for 0-back to 2-back.

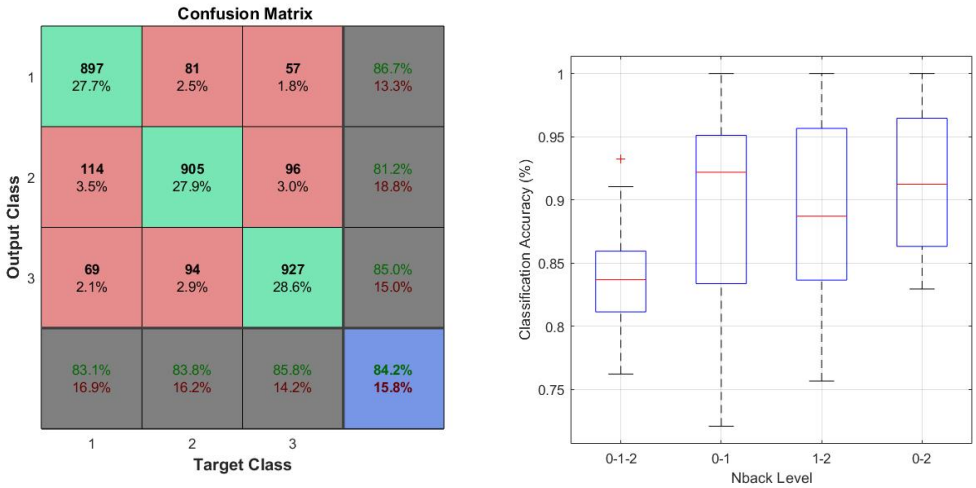

FIG. 53: Result of combining EMG and movement features with ANN classifier.

ANN consistently performed better with movement features. Using EMG and movements combined resulted in an accuracy of 84.2%, 4.1% better than LDA. The results increased by 2.1% compared to movement features and by 29.8% for EMG. For two classes the accuracy resulted in 87.5% for 0-back to 1-back, 85.5% for 1-back to 2-back and 93.2% for 0-back to 2-back.

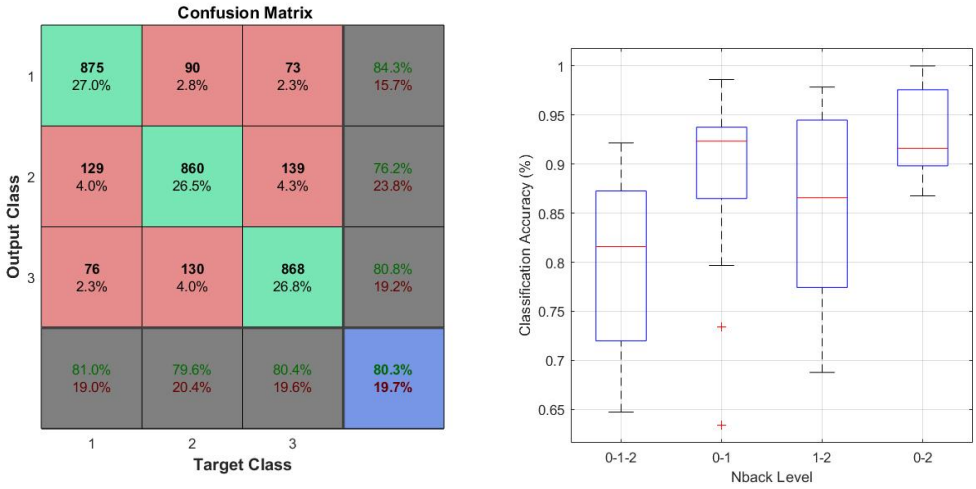

FIG. 54: Result of combining EEG and movement features with LDA classifier.

The third setup uses the movement features with the EEG frequency range between 0 - 58 Hz. Combining EEG and movement features resulted in a similar result like the EMG and movement combination. Figure 54 shows the classification accuracy for LDA. For three classes, classification resulted at 80.3% and improved by 0.2% to movement features and by 13.3% compared to EEG. The two class classifier could achieve 90.5% for 0-back to 1-back, 86.3% for 1-back to 2-back and 93.5% for 0-back to 2-back.

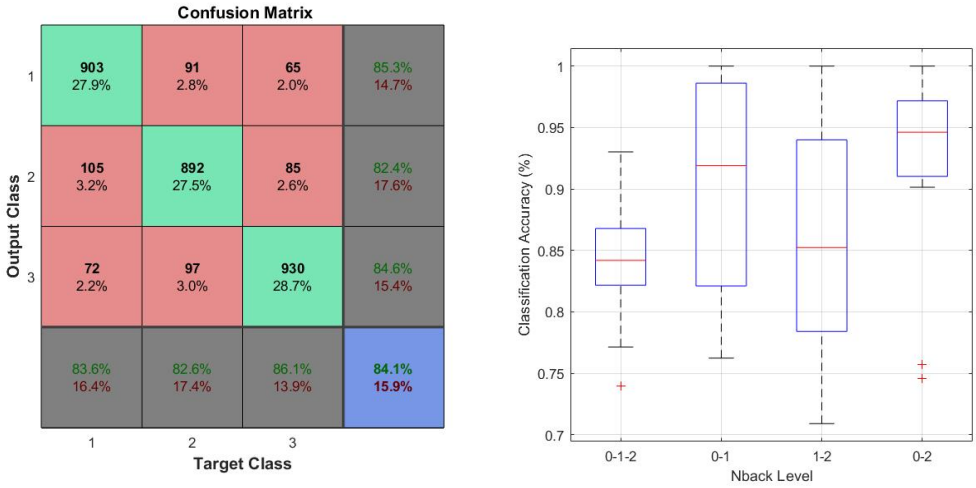

FIG. 55: Result of combining EEG and movement features with ANN classifier.

Combining EEG and movements resulted in an accuracy of 84.1%, 3.8% better than LDA. The results increased by 2% compared to movement features and by 23.8% for EEG. For two classes, the accuracy resulted in 88.2% for 0-back to 1-back, 85.6% for 1-back to 2-back and 92.9% for 0-back to 2-back.

9.1 THE BEST-PERFORMING SETUP

Lastly, all three features are combined using a bipolar derivation, WCF with the resultant controller distance as source and using the 6 highest correlations, PCA, ICA, regression denoising from 12 - 58 Hz and the hampel outlier filter. EEG and EMG features were decomposed to frequency components using Welch's method with 64 FFT points, 32 point windows and 60% overlap. Frequency bands were selected as follows: delta (0 - 4 Hz), theta (4 - 8 Hz), alpha (8 - 14 Hz), beta (14 - 30 Hz), gamma (30 - 58 Hz) and EMG (90 - 118Hz). Movement features consisting of combined controller and headset movements in three dimension and the resultant distance. All features were then combined and parsed by n-back difficulty. LDA and ANN were used as classification algorithms with 5-fold cross validation.

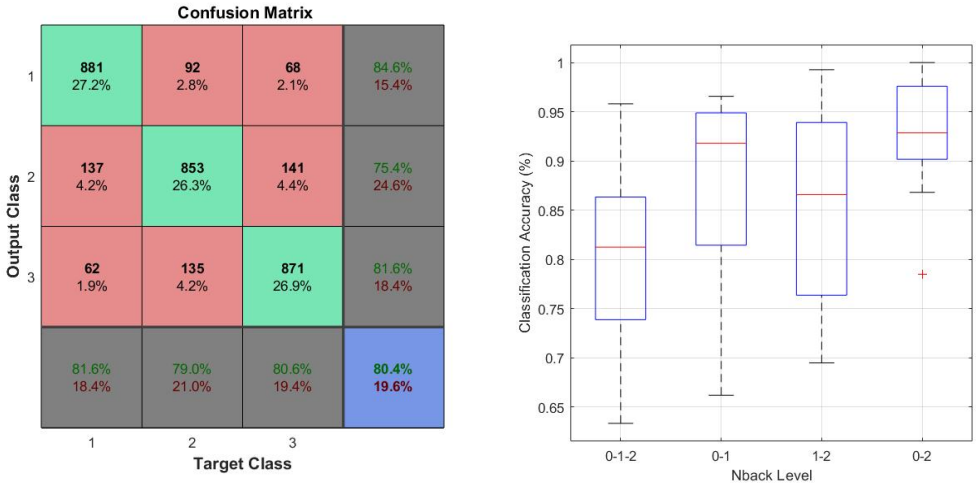

FIG. 56: Result of combining EMG, EEG and movement features with LDA classifier.

The combination of all features yielded in the highest classification accuracy for LDA with 80.4% for three classes. The classification accuracy for two classes resulted in 90.5% for 0-back to 1-back, 88.8% for 1-back to 2-back and 92.8% for 0-back to 2-back.

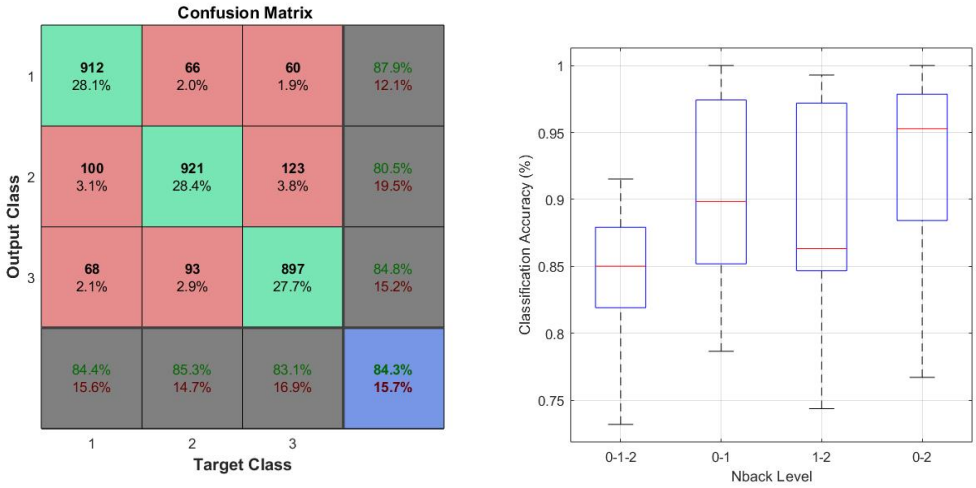

FIG. 57: Result of combining EMG, EEG and movement features with ANN classifier.

The best results of the experiment could be achieved using three classes with ANN and 84.3%. For two classes the resulted was 87.7% for 0-back to 1-back, 86%

for 1-back to 2-back and 92.9% for 0-back to 2-back.

9.2 DISCUSSION

This chapter presented feature combinations of movement, EEG and EMG, and evaluated their performance. All combinations improved the results compared to features from a single source. The best improvement could be achieved combining EMG and EEG features, which resulted in an increase of 4.1% compared to the highest classification accuracy of the combined features. The best classification accuracy for three classes could be achieved using ANN with a combination of all three features with 84.3%. The best setup for classification results for two classes was LDA with a combination of all three features with 90.5% for 0-back to 1-back, 88.8% for 1-back to 2-back and 92.8% for 0-back to 2-back. The movement features made the most significant contribution to the classification result. Combining these features with EMG, EEG or both yielded in comparable accuracy for three classes of 84.1% - 84.3% with ANN and 80.1% - 80.4% with LDA.

CHAPTER 10

MAIN CONTRIBUTIONS AND FUTURE WORK

The focus of this dissertation was to characterize and decode scalp-recorded electrophysiological activity in conjunction with kinematic activity for mental workload estimation in the context of interactive virtual reality. This chapter concludes this dissertation with the main contributions and possible future directions of this research.

10.1 MAIN CONTRIBUTIONS

The results of this dissertation demonstrate that cognitive workload during an interactive VR task can be reliably estimated via EEG. Using the traditional theta, alpha and beta EEG bands, average workload classification accuracy across reached 66.3% (chance 33.3%) for 0-back vs 1-back vs 2-back and 79.1% (chance 50%) for the extreme workload levels 0-back vs 2-back. Correlation analysis between the EEG features and controller movement revealed a high correlation, and indicates contamination of movement artifacts. In order to reduce the impact of these artifacts in EEG recordings, a new technique called the warp correlation filter (WCF), was developed that can significantly reduce the correlation between EEG and kinematic activity. The filter aligns movement features using a time-warping technique and subtracts EEG windows with similar movements from each other to compensate for the movement artifacts. The decontaminated low frequency delta band could then be added to the classification problem and improve the results to 68.3% for three classes and 79.9% for 0-back vs 2-back.

A separate correlation analysis between EEG and high-frequency EMG revealed contamination of the alpha, beta and gamma band. Regression denoising, by using

an EMG surrogate, was used to reduce theses artifacts and to enable the use of the noise-suppressed gamma band for classification. Since EMG has a positive correlation with workload, decontamination of EMG artifacts reduced the classification results slightly to 67% for three classes and 78% for 0-back vs 2-back. A standard score analysis of the raw and decontaminated EEG features revealed two pulse-decay effects that elevate EEG power in the relevant frequency bands in the beginning of each trial and at the start of each experiment, respectively. While the effect at the start of each experiment was already described in literature [44] and attributed to initial task adaption or novelty response, there has not been a description of the pulse-decay effect at the start of each trial. While the exact reasons for this pulse-decay is unclear, a temporary preparatory state of increased focus right before the task is one possible explanation.

Correlation and across-subject classification analysis highlighted the individual differences in brain activity to increased workload, which requires the development of subject specific-models. Comparison between significant feature bands before and after decontamination exhibits a reduction in significant gamma features, additionally revealing the influence of EMG activity in that frequency band. Most significant features could be found in the delta, theta and beta band. Analyzing the significance of EEG channels and the respective brain areas showed better results from electrodes located on the left brain hemisphere. The increase in significance resulted most likely because subjects used their right hands for the n-back task and the brain is processing movement of the right side on the left hemisphere. Furthermore, electrodes in the frontal brain areas showed a higher peak correlation in the theta band compared to electrodes placed at the motor cortex and parietal lobe. Increased theta activity in the frontal lobe is generally associated with increased task demand in working memory tasks [80].

Classification of the movement and EMG noise sources also revealed a modulation

by workload. Headset and controller movement features resulted in a very high classification accuracy of 82.1%, while EMG features could achieve 57% for three classes. Even though headset movements were not a part of the n-back task, they could produce a better classification accuracy (72.5%) compared to EMG and EEG. It is likely that subjects may have subconsciously performed slightly different head movements for the different workload levels. Combining these the three feature groups EEG, EMG and headset and controller movement resulted in the best possible classification accuracy with 84.3% for three classes and 92.9% for the extreme workload levels 0-back vs 2-back.

10.2 FUTURE DIRECTIONS

There are several possible future directions of this research. The most obvious is an improved experimental design for EEG-based workload studies in general. The two identified pulse-decay effects in Chapter 8 confound classification and feature analysis alike. Since the spectral power pulse-decay at the beginning of the experiment is triggered by novelty response, a longer training period could be implemented. It could be easily verified by tracking the variance of input devices like the VR controller or a computer mouse. A decline of variance over trials would indicate that the novelty effect is still prevalent. It might also be possible to forgo the use of a training period entirely and increase the experiment time. The experiment would have to be designed in a way that the first minutes of data recording could be discarded and still result in a balanced feature set. The pulse-decay effect at the beginning of each trial is harder to accommodate since it is consistent across trials over the length of the experiment. Discarding the first time windows of each trial and increasing the number of stimuli in each trial by the discarded windows is currently the most logical way to manage the impact of this effect on classification.

While the decontamination methods presented in Chapter 6 and 7 could appreciably decrease the correlation between movement and muscle tension-generated artifacts and EEG significantly, no method exists that can definitively separate EEG and EMG, occupying the same spectral frequency bands. Using the higher-frequency bands as an EMG surrogates can generate a reasonable approximation, but regression denoising only reliably eliminates EMG with a linear relationship across frequency bands. If this relationship is not linear, then residual EMG artifact will be present in the EMG suppressed signal. Further, regression decontamination should only be performed until the alpha band, since it would suppress genuine theta-band EEG relationships with the task. Therefore, the only option for definitively characterizing the relative contribution of EEG would be to paralyze muscles in the neck and face region as done by Fitzgibbon et al. [42].

There are two ways for gross-movement artifacts suppression. Either by a traditional workload study on a standard computer monitor using a keyboard or mouse as input [6, 33, 37], or as an alternative by computational modeling approaches that can estimate movement artifacts from the body movements [49]. Ultimately, for practical cognitive workload discrimination, it may not be necessary as EEG, muscle tension-related EMG and controller and headset movements contribute to cognitive workload. However, care must be taken such that EMG or movement measures do not become consciously or subconsciously conditioned for manipulating the task outcome in closed-loop scenarios.

In order to test more than the four workload levels (0-back to 3-back) that are traditionally evaluated with the n-back task, other cognitive tasks should be developed and/or evaluated. This can be done by selecting classical cognitive psychology based tasks or by alternative tasks such as video games. Examples of such tasks may include: grid location memory, digit span or mental arithmetic tasks as used by Berka et al. [44] or such as the ever popular "Tetris" by Naumann et al. [43]. It is

also prudent to increase the sample size for EEG-based workload estimation to further explore common patterns within the population, or to identify common activity within smaller subsets. Each of the 15 subjects that participated in this study showed an individual response to workload as the correlation analysis between EEG frequency bands and workload difficulty level indicated.

An alternate direction could be to exclusively focus on the analysis of of EMG and movement features, as they are more robust and easier to detect compared to EEG. Motion capture of hand or head movements can be achieved by tracking input devices like VR controller, computer mice, headsets etc., or by using motion tracking cameras. While it is known that movement reaction times are delayed with increased workload it might be worth analyzing the features in greater detail, as the combination of features increased the classification results. If reaction time would be the only information encoded in the movements, a combination would not increase the accuracy. This could be especially challenging since every subject had specific movement patterns that differentiated greatly from the others. Similar to EEG, it might be worth striving for more subjects to analyze the individual movement strategies and find common denominators.

EMG can be detected by placing electrodes on the skin near the temporalis and frontalis muscle. The application of electrolyte gel is not required since EMG measurements are not impaired by hair. The elimination of electrolyte gel increases the practicality of application and enables everyday use. While the EMG classification results in the present analysis were inferior to EEG or movement features, this dissertation only took a very small frequency band from the EEG frequency range as an EMG surrogate. Analyzing the full frequency range between 0 - 300 Hz will most likely reveal more information and a better classification accuracy.

However, as the best classification accuracy in this dissertation was achieved by a combination of features, a multi-sensor approach will most likely result in more

reliable setups. Individual measures contain most likely some degree of complimentary information for classification as other studies have shown [6]. Especially, since multiple physiological sensors like EOG, ECG, EMG etc. are common in BCI-related research, as they are used for decontamination techniques for kinematic and electrophysiological artifacts, it should not provide a challenge for researchers to synchronize these devices for feature combinations.

Ultimately, such workload estimation would be provided as closed-loop biofeedback to allow for user or task adaptation. This closed-loop scenario introduces another set of challenges as appropriate feedback and adaptation rates must be explored such that the user is effectively able to modulate the physiological activity in response to co-adapting task demands.

In any type of workload classification there is still room for improvement by using more sophisticated classification algorithms such as adaptive classifiers, matrix classifiers or deep learning techniques [81]. Especially, the popular convolutional neural networks have provided excellent performance for data-rich classification problem [82].

10.3 CONCLUSION

The results obtained in this research will pave the way for future workload estimation and EEG-based studies in VR. Virtual environments enable endless possibilities of simulating real-life situations without the necessity of leaving the controlled environment of laboratories. Particularly, the good classification results for unobtrusive passive mental state estimation will inspire researchers to similar experiment setups with increased immersiveness. The WCF technique will especially benefit researchers to compensate stereotypical, repetitive movement artifacts from contaminating EEG. The analysis of EEG features to increased workload will also help future workload experimenters. This research will encourage experimenters to plan ahead in order to accommodate for pulse-decay effects, which in turn will increase the validity of

feature analysis.

Moreover, the classification accuracy achieved in combining kinematic and electrophysiological features is at a level that can be used for practical application. As interactive, room-scale VR applications continue to become commonplace, the findings of this dissertation represent a first step toward addressing the critical issues of the integration of physiological monitoring and management of movement artifacts in VR systems.

BIBLIOGRAPHY

[1] Raja Parasuraman and Glenn F Wilson. Putting the brain to work: Neuroergonomics past, present, and future. *Human factors*, 50(3):468–474, 2008.

[2] JM Rolfe and SJE Lindsay. Flight deck environment and pilot workload: Biological measures of workload. *Applied ergonomics*, 4(4):199–206, 1973.

[3] Nathan R Bailey, Mark W Scerbo, Frederick G Freeman, Peter J Mikulka, and Lorissa A Scott. Comparison of a brain-based adaptive system and a manual adaptable system for invoking automation. *Human factors*, 48(4):693–709, 2006.

[4] William B Rouse. Adaptive aiding for human/computer control. *Human Factors*, 30(4):431–443, 1988.

[5] Walter W Wierwille and F Thomas Eggemeier. Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors*, 35(2):263–281, 1993.

[6] Maarten A Hogervorst, Anne-Marie Brouwer, and Jan BF Van Erp. Combining and comparing eeg, peripheral physiology and eye-related measures for the assessment of mental workload. *Frontiers in neuroscience*, 8:322, 2014.

[7] Edward J Haug. Feasibility study and conceptual design of a national advanced driving simulator. final report. Technical report, 1990.

[8] Benjamin Blankertz, Michael Tangermann, Carmen Vidaurre, Siamac Fazli, Claudia Sannelli, Stefan Haufe, Cecilia Maeder, Lenny E Ramsey, Irene Sturm, Gabriel Curio, et al. The berlin brain–computer interface: non-medical uses of bci technology. *Frontiers in neuroscience*, 4:198, 2010.

[9] Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain–computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.

[10] Shreya Chakrabarti. *Characterization and Decoding of Speech Representations from the Electrocorticogram*. PhD thesis, Old Dominion University, 2015.

[11] Wayne K Kirchner. Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, 55(4):352, 1958.

[12] Mark S Young, Karel A Brookhuis, Christopher D Wickens, and Peter A Hancock. State of science: mental workload in ergonomics. *Ergonomics*, 58(1):1–17, 2015.

[13] Gavriel Salvendy. *Handbook of human factors and ergonomics*. John Wiley & Sons, 2012.

[14] John G Casali and Walter W Wierwille. A comparison of rating scale, secondary-task, physiological, and primary-task workload estimation techniques in a simulated flight task emphasizing communications load. *Human Factors*, 25(6):623–641, 1983.

[15] Rebecca L Charles and Jim Nixon. Measuring mental workload using physiological measures: a systematic review. *Applied ergonomics*, 74:221–232, 2019.

[16] Andreas Henelius, Kati Hirvonen, Anu Holm, Jussi Korpela, and Kiti Muller. Mental workload classification using heart rate metrics. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1836–1839. IEEE, 2009.

[17] M De Rivecourt, MN Kuperus, WJ Post, and LJM Mulder. Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics*, 51(9):1295–1319, 2008.

[18] Stephen H Fairclough, Louise Venables, and Andrew Tattersall. The influence of task demand and learning on the psychophysiological response. *International Journal of Psychophysiology*, 56(2):171–184, 2005.

[19] Richard W Backs, H Ted Navidzadeh, and Xidong Xu. Cardiorespiratory indices of mental workload during simulated air traffic control. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 44, pages 89–92. SAGE Publications Sage CA: Los Angeles, CA, 2000.

[20] Mariel Grassmann, Elke Vlemincx, Andreas von Leupoldt, Justin M Mittelstädt, and Omer Van den Bergh. Respiratory changes in response to cognitive load: a systematic review. *Neural plasticity*, 2016, 2016.

[21] Stephen H Fairclough and Louise Venables. Prediction of subjective states from psychophysiology: A multivariate approach. *Biological psychology*, 71(1):100–110, 2006.

[22] JA Veltman and AWK Gaillard. Physiological indices of workload in a simulated flight task. *Biological psychology*, 42(3):323–342, 1996.

[23] Carolyn E Adams and Michael B Leverland. Environmental and behavioral factors that can affect blood pressure. *The Nurse practitioner*, 10(11):39–40, 1985.

[24] Glenn F Wilson. An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology*, 12(1):3–18, 2002.

[25] Dag Alnæs, Markus Handal Sneve, Thomas Espeseth, Tor Endestad, Steven Harry Pieter van de Pavert, and Bruno Laeng. Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus. *Journal of vision*, 14(4):1–1, 2014.

[26] Malcolm Brenner, E Thomas Doherty, and Thomas Shipp. Speech measures indicating workload demand. *Aviation, space, and environmental medicine*, 1994.

[27] Christian Herff, Dominic Heger, Ole Fortmann, Johannes Hennrich, Felix Putze, and Tanja Schultz. Mental workload during n-back task—quantified in the prefrontal cortex using fnirs. *Frontiers in human neuroscience*, 7:935, 2014.

[28] Lorraine Valladares. Effects of caffeine on cognitive tasks. 2009.

[29] Hasan Ayaz, Ben Willems, B Bunce, Patricia A Shewokis, Kurtulus Izzetoglu, Sehchang Hah, A Deshmukh, and Banu Onaral. Cognitive workload assessment of air traffic controllers using optical brain imaging sensors. *Advances in understanding human performance: Neuroergonomics, human factors design, and special populations*, pages 21–31, 2010.

[30] Susanne M Jaeggi, Martin Buschkuehl, Walter J Perrig, and Beat Meier. The concurrent validity of the n-back task as a working memory measure. *Memory*, 18(4):394–412, 2010.

[31] KM Miller, CC Price, MS Okun, H Montijo, and D Bowers. Is the n-back task a valid neuropsychological measure for assessing working memory? *Archives of Clinical Neuropsychology*, 24(7):711–717, 2009.

[32] Adrian M Owen, Kathryn M McMillan, Angela R Laird, and Ed Bullmore. N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping*, 25(1):46–59, 2005.

[33] Anne-Marie Brouwer, Maarten A Hogervorst, Jan BF Van Erp, Tobias Heffelaar, Patrick H Zimmerman, and Robert Oostenveld. Estimating workload using eeg spectral power and erps in the n-back task. *Journal of neural engineering*, 9(4):045008, 2012.

[34] Ying Liu, Mingtian Zhong, Chang Xi, Xinhu Jin, Xiongzhao Zhu, Shuqiao Yao, and Jinyao Yi. Event-related potentials altered in patients with borderline personality disorder during working memory tasks. *Frontiers in behavioral neuroscience*, 11:67, 2017.

[35] A Fink, RH Grabner, C Neuper, and AC Neubauer. Eeg alpha band dissociation with increasing task demands. *Cognitive brain research*, 24(2):252–259, 2005.

[36] Mirka Pesonen, Heikki Hämäläinen, and Christina M Krause. Brain oscillatory 4–30 hz responses during a visual n-back memory task with varying memory load. *Brain research*, 1138:171–177, 2007.

[37] Christian Mühl, Camille Jeunet, and Fabien Lotte. Eeg-based workload estimation across affective contexts. *Frontiers in neuroscience*, 8:114, 2014.

[38] Jeffrey B Brookings, Glenn F Wilson, and Carolyne R Swain. Psychophysiological responses to changes in workload during simulated air traffic control. *Biological psychology*, 42(3):361–377, 1996.

[39] L Finsen, K Søgaard, C Jensen, V Borg, and H Christensen. Muscle activity and cardiovascular response during computer-mouse work with and without memory demands. *Ergonomics*, 44(14):1312–1329, 2001.

[40] Bjarne Laursen, Bente Rona Jensen, Anne Helene Garde, and Anker Helms Jørgensen. Effect of mental and physical demands on muscular activity during the use of a computer mouse and a keyboard. *Scandinavian journal of work, environment & health*, pages 215–221, 2002.

[41] Danuta Roman-Liu, I Grabarek, P Bartuzi, and W Choromański. The influence of mental load on muscle tension. *Ergonomics*, 56(7):1125–1133, 2013.

[42] SP Fitzgibbon, D DeLosAngeles, TW Lewis, DMW Powers, EM Whitham, JO Willoughby, and KJ Pope. Surface laplacian of scalp electrical signals and independent component analysis resolve emg contamination of electroencephalogram. *International Journal of Psychophysiology*, 97(3):277–284, 2015.

[43] Laura Naumann, Matthias Schultze-Kraft, Sven Dähne, and Benjamin Blankertz. Prediction of difficulty levels in video games from ongoing eeg. In *International Workshop on Symbiotic Interaction*, pages 125–136. Springer, Cham, 2016.

[44] Chris Berka, Daniel J Levendowski, Michelle N Lumicao, Alan Yau, Gene Davis, Vladimir T Zivkovic, Richard E Olmstead, Patrice D Tremoulet, and Patrick L Craven. Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, 78(5):B231–B244, 2007.

[45] Ilkka Kosunen, Mikko Salminen, Simo Järvelä, Antti Ruonala, Niklas Ravaja, and Giulio Jacucci. Relaworld: neuroadaptive and immersive virtual reality meditation system. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 208–217. ACM, 2016.

[46] Liu Mingyu, Wang Jue, Yan Nan, and Yang Qin. Development of eeg biofeedback system based on virtual reality environment. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 5362–5364. IEEE, 2006.

[47] Audrey Girouard, Erin Treacy Solovey, and Robert JK Jacob. Designing a passive brain computer interface using real time classification of functional near-infrared spectroscopy. *International Journal of Autonomous and Adaptive Communications Systems*, 6(1):26–44, 2013.

[48] Thorsten O Zander, Matti Gaertner, Christian Kothe, and Roman Vilimek. Combining eye gaze input with a brain–computer interface for touchless human–computer interaction. *Intl. Journal of Human–Computer Interaction*, 27(1):38–51, 2010.

[49] Kristine L Snyder, Julia E Kline, Helen J Huang, and Daniel P Ferris. Independent component analysis of gait-related movement artifact recorded using eeg electrodes during treadmill walking. *Frontiers in human neuroscience*, 9:639, 2015.

[50] Charles S Rebert and David W Low. Differential hemispheric activation during complex visuomotor performance. *Electroencephalography and clinical Neurophysiology*, 44(6):724–734, 1978.

[51] Simone Benedetto, Marco Pedrotti, Luca Minin, Thierry Baccino, Alessandra Re, and Roberto Montanari. Driver workload and eye blink duration. *Transportation research part F: traffic psychology and behaviour*, 14(3):199–208, 2011.

[52] Tim Halverson, Justin Estepp, James Christensen, and Jason Monnin. Classifying workload with eye movements in a complex task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 56, pages 168–172. Sage Publications Sage CA: Los Angeles, CA, 2012.

[53] Frederick G Freeman, Peter J Mikulka, Lawrence J Prinzel, and Mark W Scerbo. Evaluation of an adaptive automation system using three eeg indices with a visual tracking task. *Biological psychology*, 50(1):61–76, 1999.

[54] Dennis J McFarland, Lynn M McCane, Stephen V David, and Jonathan R Wolpaw. Spatial filter selection for eeg-based communication. *Electroencephalography and clinical Neurophysiology*, 103(3):386–394, 1997.

[55] Deitrich Lehmann. Principles of spatial analysis. *Handbook of electroencephalography and clinical neurophysiology: Methods of analysis of brain electrical and magnetic signals*, 1:309–354, 1987.

[56] Ian Jolliffe. *Principal component analysis*. Springer, 2011.

[57] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[58] Christian A Kothe and Scott Makeig. Estimation of task workload from eeg data: new and current tools and perspectives. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6547–6551. IEEE, 2011.

[59] Matthias Scholz. Approaches to analyse and interpret biological profile data. 2006.

[60] Algobeans. Principal component analysis tutorial, 2016.

[61] Peter Welch. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*, 15(2):70–73, 1967.

[62] Pega Zarjam, Julien Epps, Fang Chen, and Nigel H Lovell. Estimating cognitive workload using wavelet entropy-based features during an arithmetic task. *Computers in biology and medicine*, 43(12):2186–2195, 2013.

[63] Isabelle Guyon. A scaling law for the validation-set training-set size ratio. *AT&T Bell Laboratories*, pages 1–11, 1997.

[64] Peter A Lachenbruch and M Goldstein. Discriminant analysis. *Biometrics*, pages 69–85, 1979.

[65] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[66] Simon Haykin. *Neural networks*, volume 2. Prentice hall New York, 1994.

[67] Sebastian Raschka. Linear discriminant analysis – bit by bit, 2014.

[68] Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. A review of classification algorithms for eeg-based brain–computer interfaces. *Journal of neural engineering*, 4(2):R1, 2007.

[69] Sebastian Raschka. Python machine learning.

[70] Milos Gajdos. Fun with neural networks in go, 2016.

[71] Bekir Karlik and A Vehbi Olgac. Performance analysis of various activation functions in generalized mlp architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems*, 1(4):111–122, 2011.

[72] Unity Technologies. Unity game engine, 2005.

[73] Kathryn M McMillan, Angela R Laird, Suzanne T Witt, and M Elizabeth Meyerand. Self-paced working memory: Validation of verbal variations of the n-back paradigm. *Brain research*, 1139:133–142, 2007.

[74] Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

[75] Herman Snellen. *Probebuchstaben zur bestimmung der sehschärfe*, volume 1. H. Peters, 1873.

[76] John F Golding. Predicting individual differences in motion sickness susceptibility by questionnaire. *Personality and Individual differences*, 41(2):237–248, 2006.

[77] JH Clark. The ishihara test for color blindness. *American Journal of Physiological Optics*, 1924.

[78] Byeongnam Kim, Laehyun Kim, Yun-Hee Kim, and Sun K Yoo. Cross-association analysis of eeg and emg signals according to movement intention state. *Cognitive Systems Research*, 44:1–9, 2017.

[79] Anne Collins and Etienne Koechlin. Reasoning, learning, and creativity: frontal lobe function and human decision-making. *PLoS biology*, 10(3):e1001293, 2012.

[80] Ole Jensen and Claudia D Tesche. Frontal theta activity in humans increases with memory load in a working memory task. *European journal of Neuroscience*, 15(8):1395–1399, 2002.

[81] Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005, 2018.

[82] Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U Rajendra Acharya. Deep learning for healthcare applications based on physiological signals: A review. *Computer methods and programs in biomedicine*, 2018.

VITA

Christoph Tremmel
Department of Electrical & Computer Engineering
Old Dominion University
Norfolk, VA 23529, USA

Education

Ph.D.	Old Dominion University, Norfolk, VA, USA
	Major: Biomedical Engineering, May 2015 - May 2019
Diplom-Ingenieur	Karlsruhe Institute of Technology, Karlsruhe, Germany
(M.Sc)	Major: Mechatronic Engineering, Oct 2007 - Dec 2012

Experience

Mechatronic Engineer	g.tec medical engineering, Schiedlberg, Austria
	Nov 2013 - Feb 2015
Teaching Assistant	Institute for Product Engineering, KIT, Karlsruhe, Germany
	Apr 2011 - Aug 2011
Research Assistant	Institute of Microstructure Technology, KIT, Karlsruhe, Germany
	Dec 2008 - Sep 2010

Publications

Sobreira, F., C. Tremmel, and D. J. Krusienski. "Modeling the Visual Pathway for Stimulus Optimization in Brain-Computer Interfaces." *2018 26th European Signal Processing Conference (EUSIPCO).* IEEE, 2018.

Albers, A., H. Sommer, M. Frietsch, and C. Tremmel. "Investigation of disturbances on a novel reinforcement learning based control approach" *2011 Annals of DAAAM and Proceedings.*

Tremmel, C., C. Herff and D. J. Krusienski "EEG Movement Artifact Suppression in Interactive Virtual Reality" (submitted)

Tremmel, C., C. Herff, Y. Yamani, T. Sato, K. Rechowicz and D. J. Krusienski "Estimating Cognitive Workload in an Interactive Virtual Reality Environment using EEG" (submitted)

Typeset using LaTeX.